Investigating the Evolution of mRNA: ncRNA Avoidance in *Escherichia coli*

A thesis submitted in partial fulfilment of the requirements for the degree of

Master of Science
in Cellular and Molecular Biology

at the
University of Canterbury

by
Jasper J. Perry
2018

Table of Contents

Acknowledgments

I wish to acknowledge my supervisors Paul Gardner and Anthony Poole for providing me with support and guidance throughout this project. I appreciate all the comments, criticisms and insights you've given me to further my understanding of this project. I would like to thank Ant, in particular, for bringing me into his lab at the University of Auckland and allowing me to work alongside some great people, not to mention helping me find work at the University which has allowed me to keep eating. I am very grateful!

I would also like to thank the members of the Poole Lab group who have provided me with great and weird discussions, allowed me to see things from new perspectives and offered me many interesting insights when working in the lab. I would especially like to thank Alannah who was instrumental in my understanding of lab techniques and protocols. I really appreciate how helpful and patient you were with me when I first got into the lab and how you have since continued to be. I could not have completed this work without you!

Finally, I would like to thank my family and friends for their endless support and encouragement of me. To Ellie and Saumya, thank you for sharing this experience with me. It has been great to have friends who can relate to the difficulties of student life. To my family back home in Christchurch, Mum, Dad, Max, Toby and Gen I have missed you all very much. I'm looking forward to having some down time to catch up with everyone. I hope to spend some more time with you again soon!

Abstract

It is presumed that the levels of mRNA and protein should correlate relatively strongly however this correlation is often quite poor. Two main explanations have been invoked to explain this discrepancy, messenger RNA (mRNA) secondary structure and codon usage bias, however, these explanations only account for around 40% of the total variation in expression levels. More recently a new model has been proposed that explains more of the variation in mRNA and protein levels than either codon usage or mRNA secondary structure.

The mRNA: ncRNA avoidance model, presents evidence that non-specific interactions between non-coding RNAs (ncRNAs) and mRNAs significantly impact the discrepancy between mRNA and protein abundances. The model suggests that these crosstalk interactions between mRNAs and ncRNAs impact levels of mRNA translation, consequently genes that are highly-expressed demonstrate avoidance of such interactions.

Here I present a study that investigates how highly expressed mRNAs may have evolved to avoid unintentional interactions with the abundant ncRNAs in the cell. Synonymous variants of the *araC* gene of *E. coli* were designed for increased interaction with core ncRNAs. These alterations were predicted to lead to down regulation of the *AraC* protein and subsequently impact fitness. We hypothesised that evolution of avoidance could then be driven by creating a selective pressure for high expression of *araC,* such that the affinity of the designed *araC* mRNAs for ncRNAs would be lessened to increase translation levels. The findings here demonstrate that the alterations made to the *araC* variants, which are in line with the avoidance model, have an undetectable effect on fitness in *E. coli*. Furthering our

understanding of how this phenomenon may have evolved has significant implications for the biology of RNA-RNA interaction.

Chapter 1
Introduction

The central dogma of molecular biology states that DNA is transcribed into mRNA, which is then translated into protein (DNA → mRNA → protein) (Carpenter, Ricci, Mercier, Moore, & Fitzgerald, 2014). However not all RNA codes for protein, indeed only a fraction, 1.2% of the total RNA content of the mammalian genome is protein encoding (Consortium, 2012). Similarly, prokaryotic genomes also carry large amounts of RNAs that are non-coding (Giannoukos et al., 2012; Lindgreen et al., 2014). This leaves an enormous amount of DNA that is pervasively transcribed into RNA that is not translated (Clark et al., 2011; Singh et al., 2014), in other words a large proportion of the total RNA content of a cell is non-coding. However, despite the name non-coding RNAs (ncRNAs) have many forms and functions (Herbig & Nieselt, 2011; Shabalina & Koonin, 2008), including regulation of gene expression, guiding interactions and catalysis.

The most well-known types of ncRNAs are ribosomal RNAs (rRNA) and transfer RNAs (tRNA), which are both directly involved in the decoding and translation of mRNA and constitute a large proportion of the RNA content within the cell (Giannoukos et al., 2012). Translation of mRNA is subject to extensive regulation and such regulation is often carried out by ncRNAs (Bandyra et al., 2012a; Storz, Vogel, & Wassarman, 2011; Udekwu, 2010). The regulatory activity of many ncRNAs is largely dependent on sequence complementarity with a target sequence, whereby one RNA molecule binds to another RNA molecule with a complementary sequence of base pairs to form an RNA-RNA hybrid (Mückstein et al., 2006). RNA-RNA hybrids

typically form according to Watson-Crick base-pairing rules, A:U, G:C and the wobble base pair G:U (Ananth, Goldsmith, & Yathindra, 2013) (Fig 1.2).

Theoretically protein abundance should be predictable from mRNA abundance, however mRNA levels are often an imperfect proxy for protein production (Maier, Güell, & Serrano, 2009; Tuller, Waldman, Kupiec, Ruppin, & Sherman, 2010). Two of the most widely-accepted explanations commonly invoked to explain this discrepancy are mRNA secondary structure (Mao, Liu, Liu and Tao., 2014; Tuller and Zur, 2015; Kudla, Murray, Tollervey, & Plotkin, 2009) and codon usage bias (Boël et al., 2016). However, these explanations only account for around 40% of the total variation in expression levels (Kudla et al., 2009; Plotkin & Kudla, 2011). More recently a new model has been proposed that purports to explain more of this variation in mRNA and protein levels than either codon usage or mRNA secondary structure. The mRNA: ncRNA avoidance model suggests that stochastic crosstalk interactions between mRNAs and ncRNAs significantly impact levels of mRNA translation and that highly expressed-genes are more likely to avoid such interactions (Umu, Poole, Dobson, & Gardner, 2016).

The aim of this thesis is to investigate how highly expressed mRNAs have evolved to avoid unintentional interactions with the abundant ncRNAs in the cell. To understand how protein expression is impacted via regulation of mRNAs in the following sections I will discuss the mRNA factors that are currently known to influence translation. I will then discuss the stoichiometry of molecules and macromolecules in the cell and how they impact molecular interactions. Finally, I will explain our experimental model for testing our hypothesis that natural selection drives the evolution of mRNA: ncRNA avoidance via mutations that reduce the binding affinity between such RNA molecules.

The mRNA: Protein Correlational Problem

Codon Usage Bias & Protein Expression

The expression of protein is strongly impacted by the regulation of mRNAs. mRNAs encode protein as a series of three contiguous nucleotides called a triplet codon, each of which is recognized by a tRNA. A specific region of the tRNA has complementarity with the codon on the mRNA, which form three base pairs with the mRNA, this sequence region is called the anticodon (Figure 1.1). The tRNA which physically links the mRNAs with an amino acid (Uzman, 2001) carries an amino that corresponds to its anticodon (Uzman, 2001). Each type of tRNA can be attached to only one amino acid, meaning that organisms carry many different types of tRNAs. tRNAs that carry a charged amino acid are called aminoacyl tRNAs.

During translation mRNAs and tRNAs are bound, read and processed by the ribosome at three different sites, aminoacyl (A), peptidyl (P), and exit (E) (Agrawal et al., 1996). Initiation of translation occurs when charged aminoacyl tRNAs transport the amino acid to the ribosomes A site, where the tRNA enters the ribosome. This is followed by the elongation phase whereby the tRNAs anticodon forms complementary base pairs with the codon on the mRNA at the P site allowing the amino acid attached to the tRNA to be incorporated into the growing polypeptide chain (Agrawal et al., 1996). Finally, once the amino acid has been added to the polypeptide the tRNA exits the ribosome via the E site (Sergiev et al., 2005).

While codons corresponding to the same amino acid are genetically synonymous, they are not functionally equivalent (Plotkin & Kudla, 2011). As such organisms have certain “optimal” codons that correspond to the most abundantly present iso-accepting tRNAs with the

complementary anticodons (Ikemura, 1981). This phenomenon is referred to as codon usage bias (Sharp & Li, 1987). Aminoacyl-tRNA abundance therefore impacts the rate at which codons may be translated into their amino acid counterpart (Ikemura, 1981; Plotkin & Kudla, 2011). As an example, the percentages of usage for each codon that encodes alanine in *E. coli* are as follows GCU (0.19), GCC (0.25), GCA (0.22) and GCG (0.34) (Maloy, Stewart, & Taylor, 1996). In *E. coli*, the translation rate of alanine would presumably be slowest for the GCU codon as it represents the rarest alanine codon (Sørensen, Kurland, & Pedersen, 1989), which corresponds with the number of iso-accepting tRNAs that encode the GCU codon (Dana & Tuller, 2014). In the past, it has been suggested that nucleotide changes in the third codon position, referred to as synonymous changes, have no effect on the resulting protein, given that changes at this position of the codon often specify the same amino acid, as seen with the example of alanine. Subsequently it has also been suggested that these changes should not influence cellular function, fitness or evolution. More recent studies (Boël et al., 2016; Lithwick & Margalit, 2003) however have suggested that codon usage is the most important factor in prokaryotic gene expression. The expression of heterologous transgenes in organisms provides a strong example of codon usage bias (Gustafsson, Govindarajan, & Minshull, 2004). For this reason, techniques such as codon optimisation exist that exploit this phenomenon to greatly increase protein expression. Codon optimisation assigns the most abundant codon of the host or of a selected set of genes to all instances of a given amino acid in the target sequence. In other words optimisation replaces all codons in an organism with the codons that are simply the most abundant for high expression of a heterologous transgene (Gustafsson et al., 2004; Marlatt, Spratt, & Shaw, 2010; Quax, Claassens, Söll, & van der Oost, 2015).

mRNA Secondary Structure & Protein Expression

The primary structure of RNA consists of a simple string of covalently linked A, U, G, C ribonucleotides. RNAs however do not maintain this primary structure, instead they fold onto themselves to from secondary structures via intramolecular base-pairing between the exposed ribonucleotides (Onoa & Tinoco, 2004) (Figure 1.1). This folding can also impact protein expression.

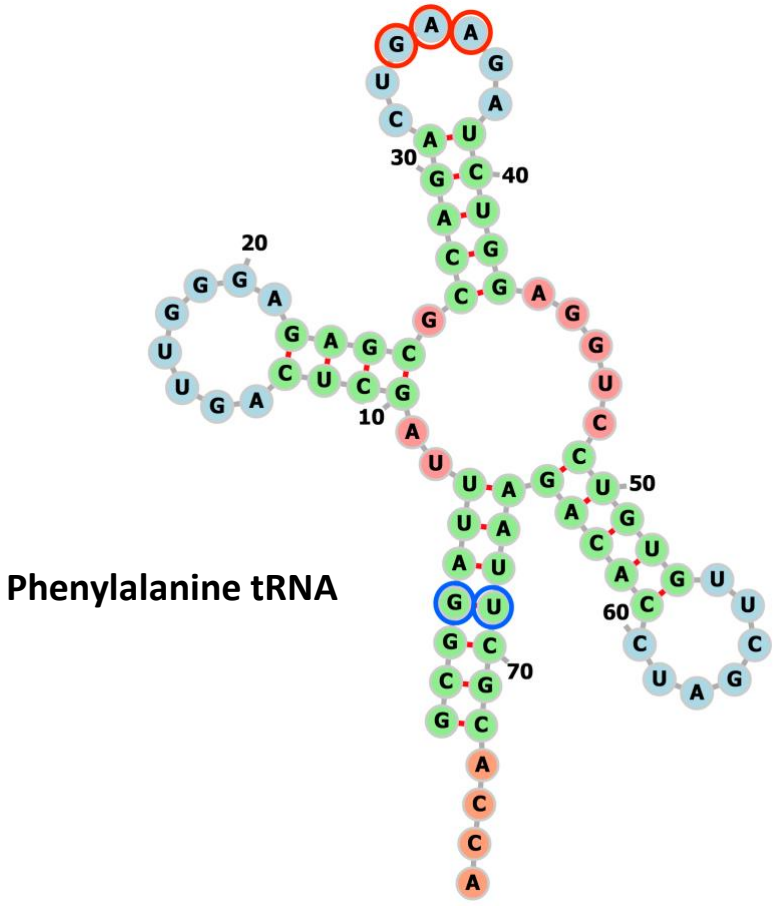


Figure 1.1. The secondary structure of the phenylalanine tRNA from *Saccharomyces cerevisiae*. Secondary structure of an RNA molecule consists of base-pairing of primary structure (i.e. sequence). This tRNA start consists of three stem loops formed by intramolecular bonding of G: C and A: U ribonucleotides, revealing the classic tRNA cloverleaf shape. The GAA anticodon is highlighted in red. The G: U wobble pairing is highlighted in blue. This image was created using the RNA drawing tool FORNA (Lorenz et al., 2011).

The capability of mRNAs to form secondary structures complicates the process of translation when secondary structure is present in the translation initiation region (TIR) of the mRNA as the strength of the interaction must be overcome by the ribosome to form a stable mRNA: ribosomal complex (Studer & Joseph, 2006). Only once a complex has been formed can translation occur. It is suggested that selection for weaker secondary structure (higher folding energy) in the 5' region of mRNAs, near the TIR, may allow faster binding of the ribosome to the transcript to initiate translation, as structures such as hairpin loops can hinder the mRNA being loaded onto the ribosome (Débarbouillé, Gabay, Schwartz, & Hall, 1982; Tuller et al., 2010; Tuller & Zur, 2015). Such an effect is universal in both prokaryotes and eukaryotes (Gu et al., 2010). This was demonstrated using synonymously variant green fluorescent protein (GFP) mRNAs, that varied randomly in their codon usage (Umu et al., 2016). Interestingly it was found that neither codon usage bias, nor the frequency of optimal codons correlated significantly with fluorescence levels of GFP. However reduced secondary structure in the 5' region was strongly correlated with higher levels of fluorescence in cultured cells (Kudla, Murray, Tollervey, & Plotkin, 2009). The strong correlation between mRNA folding and fluorescence suggests the simple mechanistic explanation that tightly folded messages obstruct translation initiation, thereby reducing protein synthesis. Indeed, during translation, ribosomes move along the mRNA to synthesise the polypeptide chain. Translation is paused wherever the ribosome encounters structure and will only continue once the base-pairing is broken (Mao, Liu, Liu, & Tao, 2014; Wen et al., 2008) (Figure 1.2). Thus, secondary structure within the message can also slow the rate of translation.

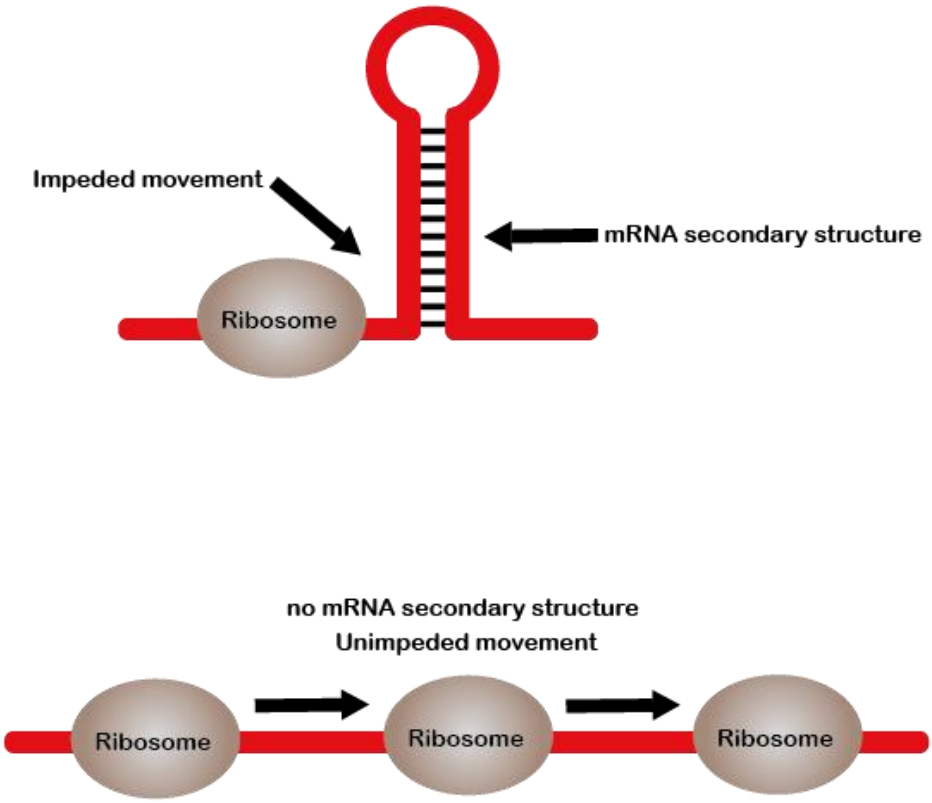


Figure 1.2: Translational inhibition by mRNA secondary structure. **A**. Stem loop secondary structuring in the mRNA slows down the movement of the ribosome along the molecule during translation. **B**. Minimal secondary structure along the molecule allows the ribosome to translate mRNA into protein without being impeded.

Synonymous changes also impact mRNA secondary structure and in turn can impact the expression level of proteins (Chamary & Hurst, 2005; Plotkin & Kudla, 2011). Indeed, secondary structure in the 5' UTR of mRNAs has been shown to limit the amount of mRNA degradation by inhibiting enzymatic activity of RNase E (Diwa, Bricker, Jain, & Belasco, 2000). Additionally, it has also been shown that poorly adapted synonymous codon choice in the 5' region of mRNAs reduces mRNA stability (Gu, Zhou, & Wilke, 2010; Plotkin & Kudla, 2011). It is suggested that selection for slow codons at the start of a gene creates a “ramp” that

prevents “traffic collisions” between ribosomes translating the same mRNA (Tuller et al., 2010).

While codon usage and mRNA secondary structure can explain a proportion of the variance associated with the poor correlation between mRNA and protein expression levels neither of these processes are able to fully explain this discrepancy. The debate as to whether secondary structure or codon usage has a greater impact on protein production has long been a point of contention. However, It is important to acknowledge that natural selection does not lead to a perfect mechanism and while neither codon usage or mRNA secondary structure correlate particularly strongly with protein levels both have been shown to impact protein expression (Kudla et al., 2009; Plotkin & Kudla, 2011; Tuller et al., 2010; Tuller & Zur, 2015).

mRNA: ncRNA Avoidance

Recently a new model has been put forward that attempts to explain the unexplained variation in mRNA and protein levels, called the mRNA: ncRNA avoidance model. This research carried out by Umu, Poole, Dobson, & Gardner (2016), demonstrates that crosstalk interactions between mRNAs and bacterial and archaeal native ncRNAs are actively avoided for highly expressed genes. Additionally, they showed that using synonymously variant GFPs with varying potentials for interaction with ncRNAs have a greater impact on protein levels than either codon usage or secondary structure. These mRNA: ncRNA interactions are the result of hybridization between native ncRNA and mRNA molecules. The model estimates that ncRNAs greatly exceed mRNAs at any site of interaction along the mRNA molecule, thus the mRNA is likely to be saturated by ncRNAs (Lindgreen et al., 2014a; Umu et al., 2016), making stochastic interactions between ncRNAs and mRNAs highly probable by chance (Lindgreen et

al., 2014b; Umu et al., 2016). Selection against mRNA: ncRNA crosstalk interactions for highly expressed transcripts in prokaryotes was calculated by computing the free energy distributions of interactions (see Chapter 2, pg 33-34) between highly conserved ncRNA and mRNA pairs. This distribution was compared to several negative controls where avoidance of interaction is unlikely. The controls included (1) di-nucleotide preserving shuffled sequences (2) homologous mRNAs from a different phylum (3) downstream regions 100 base pairs from the CDS (4) the reverse complement of the 5' end of the CDS and (5) unannotated (intergenic) genomic regions. It was found that interactions between native mRNAs and ncRNAs consistently have higher (less stable interaction) energies than expected when compared with the five different negative controls. In other words, highly-expressed messenger RNAs have a reduced capacity for hybridization with native ncRNAs and are therefore able to avoid unwanted and potentially deleterious interactions.

These explanations for the variation between mRNA and protein levels indicate the significance of mRNA regulation on the expression of protein. Codon usage, mRNA secondary structure and mRNA: ncRNA avoidance have all been demonstrated to impact protein expression. In the following section I will further explain avoidance by discussing how crowding in the cellular environment increases the potential for un-intentional molecular interactions.

Macromolecular Crowding and Stochastic Interactions

Within the cytoplasm there exists a molecular milieu that is densely packed with small and macro molecules alike. Such media are referred to as “crowded” rather than “concentrated” as no one molecule exists at particularly high concentration, but as a whole these

macromolecules occupy 5-40% of the total volume of the cell (Ellis, 2001; Mourão, Hakim, & Schnell, 2014). Crowding has been shown to impact diffusion rates (Ellis, 2001), as well as alter cellular functions like cell death, metabolism and the pathophysiology of several diseases (Lang, 2007). This phenomenon is known as macromolecular crowding or the volume exclusion phenomenon (Kuznetsova, Turoverov, & Uversky, 2014; Ralston, 1990). This effect impacts both large and small molecules, but the effect is greatest for larger molecules. The situation can be likened to a bustling airport lounge. The speed at which an adult can move through the crowd to check in will be much slower due to the difficulty in avoiding other individuals in the crowd, however a small child can more easily avoid others in the crowd allowing them to move faster (Figure 1.3 A).

The effects of crowding on biochemical reaction rates are complex because although crowding reduces diffusion (speed of movement) capacity, it increases thermodynamic activity (Ellis, 2001), due to the increased capacity for interactions between molecules. Consider a solution of identical spherical macromolecules (Figure 1.3 B). Since two molecules cannot occupy the same space, each macromolecule will exclude others from its neighbourhood in an area equal to the sum of their radii (Ralston, 1990). If we continue to add molecules the number of possible positions they can be placed becomes increasingly limited, increasing the potential for interaction.

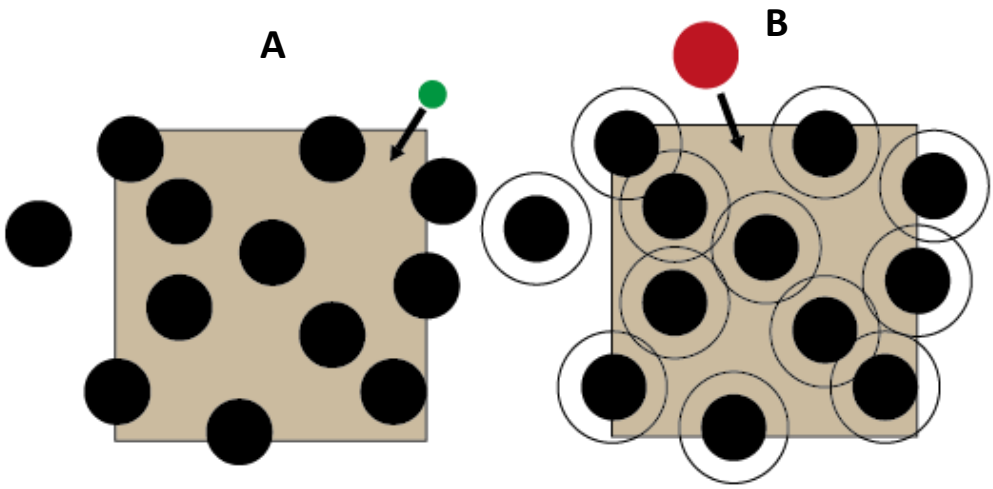


Figure 1.3: The volume exclusion phenomenon. **A**) The movement of the green molecule through the cellular environment is less inhibited due to its smaller size. **B**) The movement of the red molecule has greater difficulty moving through the cellular environment as large densely packed macromolecules hinder its ability to easily move due to its larger size. The black rings surrounding each molecule represent that molecules exclusive neighbourhood.

While binding affinity is a key aspect of molecular interactions (Kastritis & Bonvin, 2013), macromolecular crowding can lead to non-specific interactions between molecules. Protein misinteractions are directly influenced by the volume exclusion caused by macromolecular crowding (Kuznetsova et al., 2014). These misinteractions refer to non-functional and typically non-specific protein-protein interactions that occur when random protein molecules encounter one another (Yang, Liao, Zhuang, & Zhang, 2012). Protein misinteractions are a frequent occurrence in the cell for two main reasons. The first is because at any one time many proteins co-exist in close proximity, meaning the opportunity for unintentional or even deleterious interactions is increased. As proteins often only interact with a few specific

protein partners (Qian, He, Chan, Xu, & Zhang, 2011) and given that the total number of proteins that are co-expressed can be in the millions for a single *E. coli* cell (Milo, 2013), the total concentration of specific protein partners is far outweighed by the concentration of non-specific partners. The second is that despite functionally specific interactions typically being stronger than misinteractions the difference in binding energy is only moderate (Yang et al., 2012). Thus, any potential non-specific interactions can result in a strong protein complex. This phenomenon led to the development of the protein misinteraction avoidance hypothesis (Yang et al., 2012) which states that the number of potential deleterious protein misinteractions increases with concentration, and thus highly expressed proteins are under a strong selective pressure to avoid such interactions (Zhang & Yang, 2015). This hypothesis closely resembles the mRNA: ncRNA avoidance model, in that highly expressed genes demonstrate a signal for avoidance of un-intended interactions with ncRNAs.

The effect of molecular crowding has also been shown to have an influence on the folding of RNA structures such as ribozymes and other RNA molecules (Dupuis, Holmstrom, & Nesbitt, 2014; Kilburn, Roh, Guo, Briber, & Woodson, 2010; Lee, Kilburn, Behrouzi, Briber, & Woodson, 2015). The impact of crowding on interactions between RNA molecules is made evident by the mRNA: ncRNA avoidance model. At the beginning of this chapter I briefly discussed how both prokaryotic and eukaryotic genomes pervasively transcribe large amounts of their DNA into RNA which does not code for protein (ncRNA). Therefore, throughout this thesis I operate under the assumption that ncRNAs greatly exceed mRNAs. Based on the degradation rates of mRNAs compared with stable RNAs (rRNA and tRNA) (Deutscher, 2006) in addition with other previous RNA-seq (Wang, Gerstein, & Snyder, 2009) data (Giannoukos et al., 2012; Lindgreen et al., 2014) the assumption that ncRNA levels exceed those of mRNAs is not biologically

unreasonable. In an analysis of over 400 publicly available bacterial and archaeal RNA-seq datasets (Lindgreen et al., 2014) it was found that of the 922 RNAs of unknown function (RUFs) identified over half (568) were among the most highly abundant transcripts. Further to this a transcriptome analysis of bacterial communities constructed RNA-seq libraries from total RNA and rRNA-depleted samples. Without depletion, it was found that >98% of all mapped reads aligned to rRNA (Giannoukos et al., 2012). Consequently, this assumption predicts that any potential mRNA interaction regions are likely to be saturated by ncRNAs in the surrounding cellular environment.

In summary interactions between mRNAs and ncRNAs are common in both eukaryotes and bacteria. The stochasticity of these interactions between mRNAs and ncRNAs is in part due to their stoichiometry, in that ncRNAs are highly abundant, which subsequently increases the potential for interaction. Previous research has shown that highly expressed mRNAs across prokaryotes demonstrate a signal for avoiding of interactions with ncRNAs. In addition, designing mRNAs for varying interaction potentials with ncRNAs significantly impacts protein expression.

Objectives and Experimental Design

The aim of this study was to demonstrate how the mRNA: ncRNA avoidance signal may have evolved. We suspected that avoidance evolved in response to deleterious interactions between highly expressed mRNAs and ncRNAs that inhibit the translation of important genes. The findings of Umu et al. (2016) clearly demonstrated that an avoidance signal exists that works to reduce mRNA: ncRNA interactions. We hypothesised that synonymous mutations

occur within the mRNA transcript to reduce the affinity of these types RNA: RNA interactions and ensure efficient protein expression. To test this hypothesis, we created synonymous variants of the arabinose metabolism gene, *araC,* from *E. coli,* where the potential for stochastic interactions between the *araC* transcript and native ncRNAs in *E. coli* was increased. Thus, we potentially hinder translation via crosstalk with ncRNAs. By placing populations of *E. coli* carrying this gene variant under a selective pressure where arabinose was the sole carbon source, and serially passaging these lines for many generations we aimed to show that *araC* expression will gradually increase. Given the essentiality of the selected gene in this environment, regarding the organism's overall fitness, an RNA-RNA interaction that inhibits expression of the gene is predicted to be detrimental. Thus, we proposed that the selective pressure on placed *E. coli* populations would reduce ncRNA interactions with the *araC* mRNA, improve translation efficiency and increase the overall fitness of the population, leading to the evolution of avoidance. The advantage of this approach in comparison to previous research by Umu et al, (2016) is that we are testing the gene in a native context, allowing us to test the evolution of avoidance in a true biological system.

Experimental Evolution

In this research, we aimed to test the evolutionary emergence of mRNA: ncRNA avoidance but inferring evolutionary origins can be a very difficult task. However, one of the best methodologies for assessing evolutionary events is through experimental evolution (Lenski, Rose, Simpson, & Tadler, 1991). Experimental evolution is defined as the study of evolutionary changes occurring in experimental populations as a consequence of conditions imposed by the experimenter (Kawecki et al., 2012). Experimental evolution has been applied to a number of questions in biology, mainly adaptation, trade-offs and constraints, population

genetics and evolutionary theory (Kawecki et al., 2012). Evolution experiments often utilise bacteria with fast generation times and large population sizes allowing researchers to view evolution on a feasible timescale. The use of bacteria also allows for populations to be stored cryogenically and later revived to compete ancestors against descendants (Wiser, Ribeck, & Lenski, 2013). This work was pioneered by Lenski, Rose, Simpson, & Tadler (1991) who have now run the long-term evolution experiment for more than 50,000 generations. The experiment has been run using 12 identical populations of *E. coli* cultured in minimal glucose media (Barrick et al., 2009). Over the course of this experiment these lines have accumulated hundreds of mutations. One notable mutation occurred in a single line of *E. coli* which evolved citrate metabolism in oxygenated environments where it was previously unable to do so (Blount, Borland, & Lenski, 2008). The versatility of experimental evolution as a research tool is apparent in its ability to test predictions from evolutionary theory. Such studies have demonstrated for example that bacteria can evolve a new phenotypic switch (Beaumont, Gallie, Kost, Ferguson, & Rainey, 2009), also known as bet hedging, that natural selection may favour male traits that directly reduce the fitness of their mates (Rice, 1996), and that reproductive isolation can occur as a result of divergent selection in different environments (Dodd, 1989).

The Arabinose Operon

When cultured in lab *E. coli* grown on a medium containing multiple carbon sources will often preferentially metabolise only one sugar at a time. This process is known as carbon catabolite repression (Stülke & Hillen, 1999). A classic example of this is seen in *E. coli* grown on media containing both glucose and lactose. *E. coli* will firstly consume all the glucose in the media before switching to lactose (Stülke & Hillen, 1999). In addition to glucose and lactose there

are many other sugars that *E. coli* can utilise as a carbon source. In the absence of a preferred energy source *E. coli* will utilise L-arabinose, a five-carbon sugar, as a source of carbon and energy (Desai & Rao, 2010; Schleif, 2010).

The *araC* gene of the *E. coli* arabinose operon was selected as our gene of interest given its extensive history in the study of molecular systems, as well as the relative ease with which it can be assayed. Four genes *araA*, *araB, araC* and *araD* have been identified as being required for the uptake and conversion of L-arabinose to D-xyulose-5-phosphate, following which D-xyulose-5-phosphate enters the pentose phosphate pathway (Schleif, 2010). *araA* converts arabinose to L-ribulose which is subsequently phosphorylated by *araB* and converted from L-ribulose-phosphate to D-xyulose-phosphate by *araD* (Englesberg, 1961). *araC* regulates expression of its own synthesis as well as the other genes of the arabinose operon (Schleif, 2000). The arabinose operon was one of the first gene expression systems found to use positive regulation (Schleif, 2000). In the presence of arabinose *araC* stimulates expression of mRNA transcripts from promoter *BAD*. The P_{BAD} promoter of *E. coli* exhibits an all-or-nothing induction of expression, also referred to as autocatalytic gene expression (Siegele & Hu, 1997). This means that populations that are grown under conditions of subsaturating levels of inducer (arabinose) will have some cells in the population being induced whereas the other proportion will not be or will be induced at very low levels (Siegele & Hu, 1997).. When sufficient arabinose is present, the *araC* gene product remains bound by arabinose (Figure 1.4 A) allowing RNA polymerase (RNAP) to bind to the promoter region and transcribe the operons structural genes (Figure 1.4 B). Conversely the *araC* protein will also repress the expression of transcripts in the absence of arabinose by forming a dimer and binding to two regions along the operon, the initiator region *araI* and the operator $araO_2$, forming a loop

structure in the DNA (Schleif, 2010) (Figure 1.4 C) that prevents RNAP from binding to the promoter.

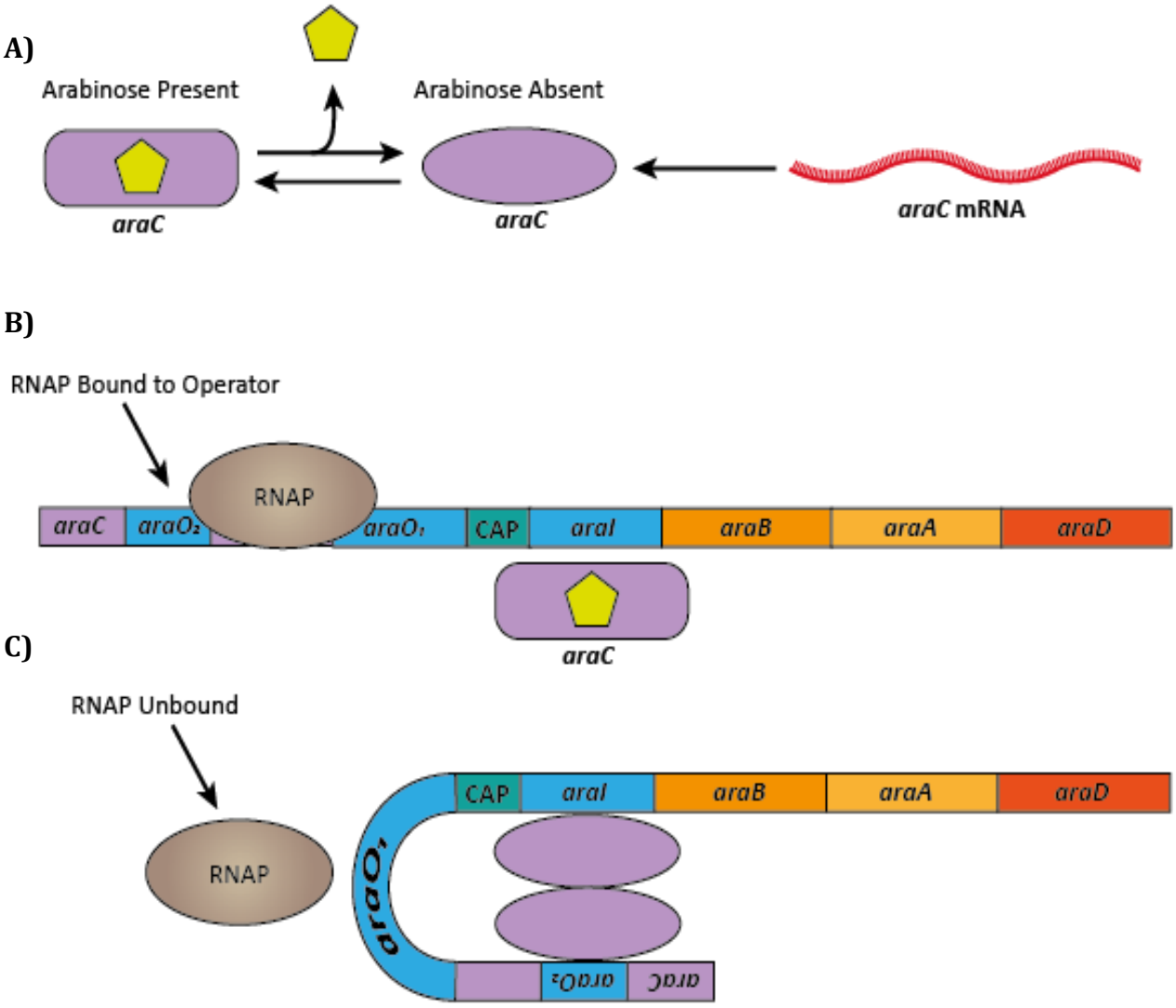


Figure 1.4: The arabinose operon **A**) The structural conformation of *araC* changes depending on whether arabinose is present or absent in the media. **B)** Binding of arabinose to araC prevents dimer formation, allowing RNA polymerase (RNAP) to bind to the operator. **C)** When *araC* is unbound two molecules come together to form a dimer that binds to the initiator region *araI* and operator *araO*$_2$ forming a loop structure, that inhibits RNAP binding and prevents translation.

Potential Adaptive Responses to Selection

While we hypothesised that unfavourable RNA-RNA interactions would be avoided via changes to the gene at the nucleotide level, there were at least three potential adaptive responses which we hypothesised that *E. coli* may have exhibited in response to imposing a selective pressure to increase protein expression:

i) In line with our hypothesis, evolution of an *araC* variant expressing lines may have resulted in crosstalk being reduced via changes at the nucleotide level of the *araC* gene. However sequences changes at this level could also be the result of altered codon usage (Quax et al., 2015) or reductions in mRNA secondary structure, both of which could increase the level of translation (Kudla et al., 2009).

ii) Any effect of crosstalk interactions could be abrogated via a promoter mutation in the arabinose operon that increases *araC* mRNA transcription (Tamai, Belyaeva, Busby, & Tsuchiya, 2000). With a greater number of expressed *araC* transcripts fewer mRNAs would may be impacted by crosstalk interactions ncRNAs.

iii) Amplification of the *araC* locus might limit the impact of crosstalk via increased expression from additional copies of the *araC* gene. Having two or more copies of *araC* may increase the amount of mRNA available for protein synthesis (Qian & Zhang, 2014).

As a corollary to the adaptive responses outlined above, if the change in growth rate is based on amplification of mRNA and, as we've established, ncRNAs are present in much greater abundance, then amplification would need to be extremely high to enable some mRNAs to slip through the "crowd" of ncRNAs, freeing them up to be translated into protein.

Chapter 2
Bioinformatics of Variant araC Design

The research carried out by Umu et al. (2016) evaluated the significance of unfavourable interactions between RNAs on protein expression. They showed that RNA-RNA crosstalk is selectively disadvantageous, and can therefore be detectable as an 'avoidance signal'. These RNA-RNA interactions negatively impact protein expression by inhibiting translation and thus appear to be underrepresented for highly expressed genes.

The premise behind this research was to, using a bioinformatics approach, create synonymous variants of the *araC* gene in *E. coli* such that the potential for stochastic interactions between the designed *araC's* transcript and native ncRNAs is increased. The *araC* variants were designed to weakly bind ncRNAs to allow natural selection to more easily release the target molecule from inhibition. It is important to avoid completely silencing the *araC* gene as this would be very detrimental for the organism under conditions where arabinose is the sole carbon source. This weak interaction makes it easier for *E. coli* to produce changes in the third codon positions, as substitution mutations at this position often result in a synonymous change (Bofkin & Goldman, 2007). As avoidance was our dependent variable, extraneous variables such as codon bias, mRNA secondary structure and G+C content needed to be controlled to ensure they have no effect on expression of *araC* in *E. coli*. To achieve this, we relied on the help of bioinformatics tools.

To understand the need for such tools in the process of designing the gene variants optimised for ncRNA interaction I will firstly describe the functions of well-known ncRNAs. Secondly, I

will discuss current issues with RNA structure and interaction prediction as well as describe the tools used for designing our gene variants and how they operate. Finally, I will describe the methodology used to create the synonymously variant *araC* mRNAs designed for high interaction with ncRNAs.

Regulatory Functions of ncRNAs

The discovery of ncRNAs and their functions is not a trivial process (Bao, Cervantes Cervantes, Zhong, & Wang, 2012; Herbig & Nieselt, 2011). However, despite the difficulty in discovery and classification several highly important ncRNAs have been well-characterized. Micro RNAs (miRNAs) and small interfering RNAs (siRNAs) in eukaryotes (Cloonan, 2015; Gomes, Nolasco, & Soares, 2013) are two well-known classes of ncRNA that regulate gene expression by base-pairing with a target sequence which typically results in down regulation of the target mRNA, otherwise known as RNA interference (RNAi). These short, 20 – 30 nucleotide RNA sequences bind to mRNA molecules, resulting in translational inhibition and or mRNA degradation (Carthew & Sontheimer, 2009). Both miRNAs and siRNAs associate with a class of proteins known as Argonautes which enable them to carry out gene silencing (Carthew & Sontheimer, 2009). miRNAs typically form RNA hybrids at specific "seed' regions with perfect complementarity to the target (Bandyra et al., 2012b). These seed regions are usually found in the 2^{nd} to 7^{th} nucleotide along the miRNA, and their target binding region typically forms in the 3' untranslated region (UTR) of the target mRNA (Lewis, Burge, & Bartel, 2005).

Similarly, bacterial small RNAs (sRNAs) utilise RNA-RNA interactions to regulate gene expression in prokaryotes (Gottesman & Storz, 2011). In many enteric bacteria, these sRNAs

consist of three main structures which I will now describe. (1) Secondary structure at the 3' to activate Rho-independent transcription termination and protect against exonuclease activity. Rho is a transcription factor that is often required for termination of transcription (Banerjee, Chalissery, Bandey, & Sen, 2006). However secondary structure in the 3' region of the sRNA facilitates transcription termination without the need for Rho. This secondary structure is followed by a poly-U stretch of nucleotides that destabilises the RNA-DNA duplex causing RNA polymerase (RNAP) to fall off, terminating transcription (Figure 2.1). (2) an Hfq binding site. Hfq is an RNA binding protein that is a common feature of bacterial lineages. This protein plays a key role in the regulation of gene expression by facilitating the pairing of sRNAs with their target mRNAs (Vogel & Luisi, 2011). Hfq facilitates suppression of translation by aiding the cognate sRNA to bind to the 5' end mRNA which hinders ribosomal access for translation. Alternatively, Hfq can also increase translation by guiding the cognate sRNA to the 5' region of the mRNA to disrupt secondary structure that is inhibiting ribosomal access. In addition, Hfq can also protect sRNAs from degradation by ribonucleases or present sRNAs to promote mRNA cleavage. Conversely Hfq may also facilitate RNA turnover by making the 3' of the mRNA accessible for degradation (Vogel & Luisi, 2011). And (3) a highly-conserved region that is utilized for target binding (Bandyra et al., 2012; Storz et al., 2011) analogous to the “seed” regions seen in eukaryotic miRNAs (Figure 2.1) (Lewis, Burge, & Bartel, 2005). (Vogel & Luisi, 2011). sRNAs range from 50-250 nucleotides in length (Vogel & Wagner, 2007) however, despite the potential for 100s or more base pairings with a cognate mRNA, sRNAs initially bind very quickly and with high affinity using only a few nucleotides that are exposed in stem loops (Gottesman & Storz, 2011, Storz, Vogel, & Wassarman, 2011; Waters & Storz, 2009) (Fig 1.2). Following this initial interaction additional base pairs can form between the sRNA and the target mRNA (Storz et al., 2011; Waters & Storz, 2009). The required binding

region in sRNAs is typically a single stranded stretch of nucleotides which suggests complex secondary structure is not important for binding (Figure 2.1) (Storz et al., 2011). The accuracy and efficiency of these "seed" regions was emphasised in an experiment wherein the conserved binding region of *Salmonella* RybB, a bacterial sRNA, was fused to unrelated scaffold RNAs. It was demonstrated that this seed region alone was sufficient to guide target recognition (Papenfort, Bouvier, Mika, Sharma, & Vogel, 2010).

The regulatory activity of ncRNAs is clearly dependent on several factors. However, this dependency is particularly strong regarding the "seed" region which can facilitate target recognition without the aid of the additional factors. This highlights the impact that stochastic RNA-RNA interactions can have on the regulation of gene expression.

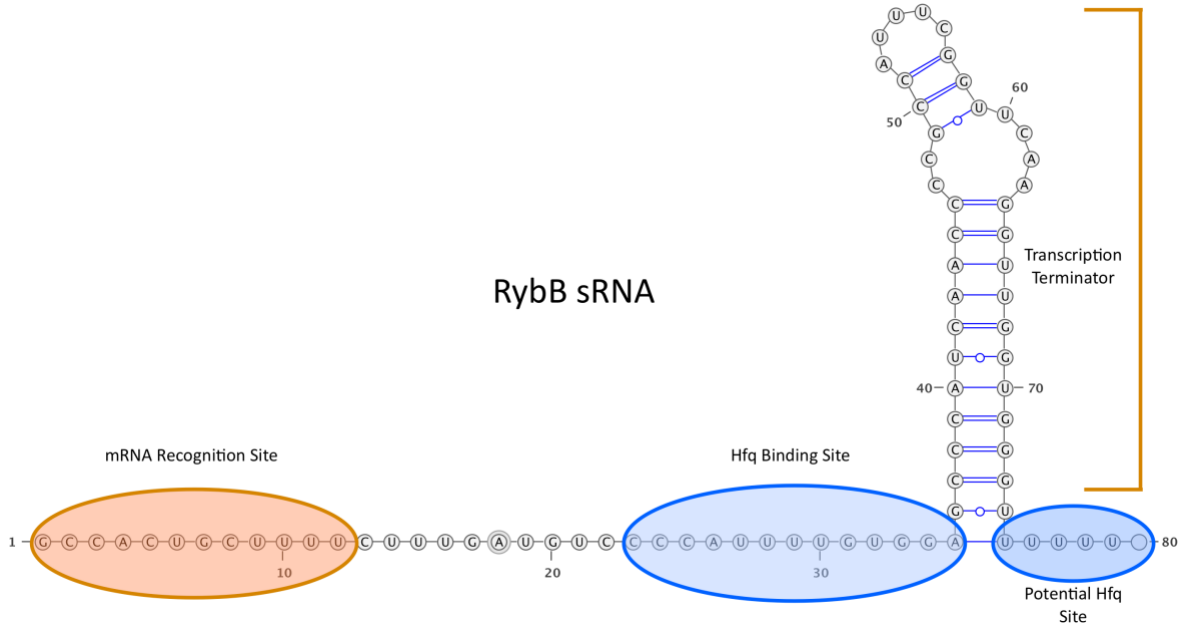


Figure 2.1. The domain structure of RybB small bacterial RNA (sRNA) from *Salmonella enterica*. The figure shows the Hfq binding site as well as a potential Hfq binding site (blue ellipses). The 'seed' region of the sRNA (orange ellipses) for mRNA recognition is also shown. The transcription terminator is the hairpin structure located at the 3' end of the sRNA

indicated by the orange line. The secondary structure in conjunction with the poly-U stretch of nucleotides destabilise RNA-DNA duplexes which causes RNAP to fall off. This figure was created using the RNA drawing tool VARNA (Darty, Denise, & Ponty, 2009).

Intramolecular and Intermolecular RNA-RNA Interactions

It has been estimated that for any one RNA molecule there are 1.8^n possible folding conformations, where *n* is the number of nucleotides. This means for a 100 ribonucleotide long RNA there are approximately 3.37×10^{25} possible conformations (Michael Zuker & Sankoff, 1984). Secondary structures are often conserved and thus allow comparisons between ncRNAs. To further emphasise this the Rfam database (Gardner et al., 2009) carries a collection of ncRNA structures and other RNA sequence families represented by covariance models (CMs) and multiple sequence alignments (MSAs). The role of structure is critical in the function of RNA molecules, however RNA secondary structure is difficult to determine experimentally (Lorenz, Wolfinger, Tanzer, & Hofacker, 2016). For this reason, many algorithms have been developed for predicting stable RNA secondary structures.

In our calculations of RNA secondary structure, we use RNAfold. RNAfold is a RNA secondary structure prediction program that calculates the minimum free energy (MFE) of intramolecular base pairs (Lorenz et al., 2011). MFE methods aim to identify the optimal folding conformation of RNAs via a dynamic programming approach (Licon, Taufer, Leung, & Johnson, 2010), whereby complex problems can be broken down into a series of sub-problems (Mückstein et al, 2006). In the case of RNA these dynamic programs work in two stages. In the first stage, minimum folding energies are computed and stored for all fragments

of the RNA sequence, this process may start with 5-nucleotide fragments and build up to larger fragments in a recursive manner. The second stage computes a minimum energy structure by searching through the matrix of stored energies (Zuker, 1989). Additionally thermodynamic parameters determined by nearest neighbour energy models (Zhang, 2016) are used to increase the number of stacked base-pairs in a predicted stable RNA with the lowest energy, where higher energy would indicate an unstable or unstructured RNA. The underlying assumptions behind nearest neighbour energy models are (1) that the stability of an RNA structure depend on the sequence of the structure and the sequence of the adjacent stacking pairs and (2) that the total stability of a structure is the sum of the stability of each pairing (Westhof & Fritsch, 2000) (Figure 2.2). As such the total free energy of an entire stacking region is contributed to by both base pair stacking and intramolecular bonding between nucleotides.

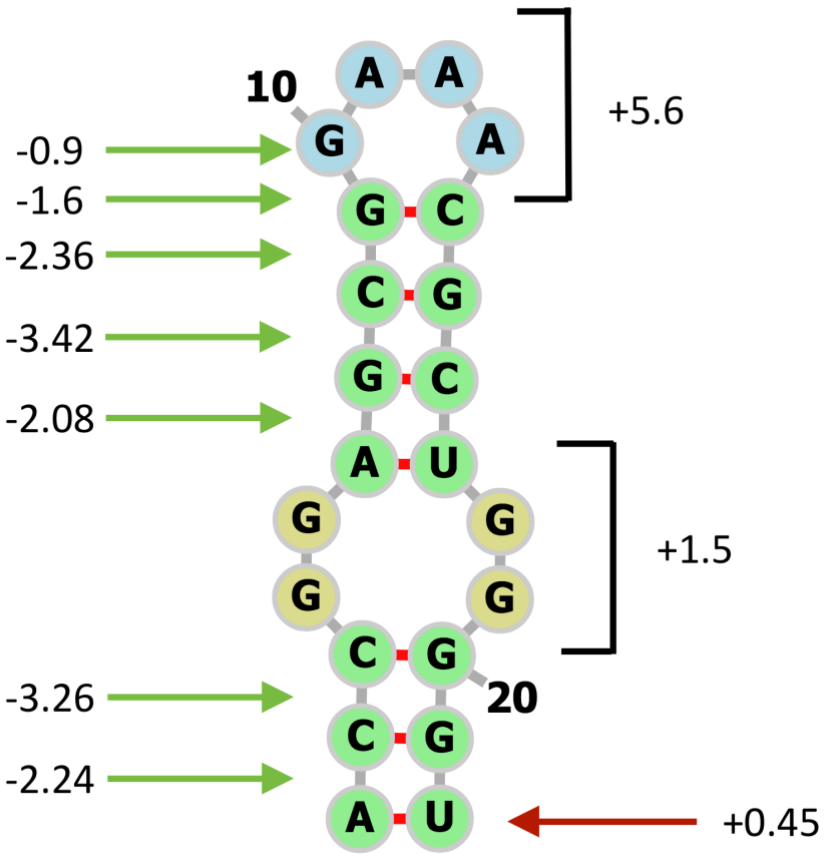


Figure 2.2. The nearest neighbour calculation for an RNA secondary structure. All calculations are in kcal/mol. Green arrows indicate the energy gained from stacking base pairs. The internal loops incur a penalty where no base-pairing occurs and energy is lost. The red arrow indicates where additional penalties are given for helices ending with A: U pairs. The total sum of this structure is -8.3kcal/mol. This structure was redrawn from Andronescu et al, (2014) using the RNA drawing tool FORNA (Lorenz et al., 2011).

Turning to intermolecular base-pairing of RNAs, given an RNA query sequence and a set of potential target RNAs, finding the correct pairing target for the query sequence can be difficult (Salari, Backofen, & Sahinalp, 2010). The term "correct" here means the RNA target with which the cognate RNA has the greatest binding affinity. It is important to note that while many RNAs are predicted to have specific partners their range of interaction is often far reaching, albeit usually less effective (Hausser & Zavolan, 2014). This presents two hypotheses: the "Few Targets" hypothesis and the "Many Targets" hypothesis. The "Few Targets" hypothesis postulates that the vast majority of weakly repressed targets are noise and that only a few strongly repressed will have any phenotypic effect. Conversely the "Many Targets" hypothesis postulates that the repression of many targets is the result of a system wide effort to stabilise gene regulatory networks (Zhao, Shen, Tang, & Wu, 2017). For this reason, predicting RNA-RNA interaction also presents a significant problem for computational biology.

Following secondary structure prediction MFE methods were also extended to RNA-RNA interaction prediction tools. MFE algorithms that predict RNA-RNA interactions can be understood as the sum of two energy values (1) the energy required to open a binding site,

or unfold the RNA, otherwise known as the RNAs accessibility and (2) the energy or stability that is gained from hybridization, or base-pairing of two RNA molecules (Mückstein et al., 2006a). RNAup is an RNA-RNA interaction prediction program that calculates the MFE thermodynamics of RNA binding (Mückstein et al., 2006a). Research has shown that the overall performances scores were highest in sensitivity and precision for RNA interaction prediction algorithms that utilise free energy minimization based algorithms by incorporating measures of binding region accessibility (Umu & Gardner, 2017). As such MFE methods constitute the majority of RNA interaction prediction tools (Umu & Gardner, 2017). In the same assessment comparing the performance of RNA-RNA interaction prediction tools RNAup was demonstrated to be the most precise, giving the highest rate of true positives compared with other prediction tools (Umu & Gardner, 2016). For this reason, we use RNAup in our calculations of RNA-RNA interactions (avoidance).

Codon Adaptation Index (CAI) Calculator

Codon adaptation values for *E. coli* B. strain REL606 were determined based on codon distribution patterns from the entire set of protein coding mRNAs developed by Umu et al, (2016) using Biopython libraries (version 1.6) (Cock et al., 2009). The adaptiveness of each codon was calculated across the whole of each *araC* variant mRNA using the codon adaptation index (CAI) (Sharp & Li, 1987).

During translation 61 codons (and three stop codons) specify 20 amino acids. The genetic code is therefore referred to as redundant. This redundancy is what leads to the phenomenon of codon usage bias (see Chapter 1, pg. 4-5). Genes that are more highly expressed

demonstrate both fast and accurate expression, as the codons that comprise the sequence of these genes are better adapted to the tRNA pool and thus will be translated more efficiently than lesser-adapted codons (Ikemura, 1981). Understanding which codons will lead to more highly expressed genes is important for protein production in bioengineering. The CAI allows us to determine which codons are most highly utilised in organisms by assessing the relative merits of each codon in a gene. In doing so a score is calculated based on the frequency of use of all the codons in that gene. This is achieved by using a reference set of highly expressed genes from the organism in question.

There are several assumptions the Codon Adaptation Index operates under, (1) the pattern of codon usage within a gene is largely determined by natural selection and mutation, (2) selection appears to occur via translational efficiency, so that synonymous codon usage in highly expressed genes is under the strongest selective constraints and (3) in E.coli and yeast, very highly expressed genes appear to have the greatest degree of synonymous codon bias (Sharp & Li, 1987). These assumptions ultimately can be broken down to mean that highly expressed genes can reveal (i) which alternative codon encoding an amino acid is most efficient for translation and (ii) the extent to which other codons are disadvantageous (Sharp & Li, 1987).

When forming a CAI, the first step is to construct a reference table of relevant synonymous codon usage (RSCU) values from highly expressed genes of the organism in question. An RSCU value is simply the observed frequency of that codon divide by the frequency that would be expected under an assumption of equal codon usage for an amino acid. The relative adaptiveness of a codon is the frequency of use of that codon, compared with the frequency

of use of the optimal codon for that amino acid (Sharp & Li, 1987). The formula for this calculation is as follows:

$$w_{ij} = RSCU_{ij} / RSCU_{imax} = X_{ij} / X_{imax}$$

Where w_{ij} is the adaptiveness of a codon and $RSCU_{imax}$ and X_{imax} are the RSCU and X values for the most frequently used codon for the ith amino acid.

The CAI metric defines how well mRNAs are optimised for codon bias (Sharp & Li, 1987). CAI values range from 0 to 1, with higher values indicating a higher proportion of the most adaptive codons (Sharp & Li, 1987). In short CAI measures the deviation of any given protein coding gene sequence from a reference set of highly expressed genes (Sharp & Li, 1987) to determine which codons are most adaptive (translated more efficiently).

Materials and Methods

Here I summarize the data sources, materials and methods implemented in the design of our *araC* gene variants. The computational methods used in this design process were performed in Python (version 2.7.12) (Cock et al., 2009) or using Bash shell scripts. All bioinformatics tools used and their versions are cited. The scripts used to carry out the design and generation of *araC* variants can be found on GitHub (https://github.com/UCanCompBio/Avoidance).

The annotation file for *E. coli* REL606 was obtained from the National Center for Biotechnology Information (NCBI) website (https://www.ncbi.nlm.nih.gov/). All ncRNAs

including rRNAs, tRNAs and other well-characterized ncRNAs: RNaseP (Altman, 2011), SRP RNA (Bradshaw & Walter, 2007) tmRNA (Keiler & Ramadoss, 2011) and 6S RNA (Steuten et al., 2014) were pulled from the *E.coli* REL606 genome using the annotation program Artemis (version 10.2) The entire set of core mRNAs was obtained through previous work by Umu et al., (2016). The 114 evolutionarily conserved (core) mRNAs were obtained from PhyEco (Wu, Wei, Liu, Li, & Rayner, 2011). PhyEco markers are based on a set profile of hidden Markov models (HMMs) (Krogh, Brown, Mian, Sjölander, & Haussler, 1994) that correspond to highly conserved bacterial protein coding genes, including ribosomal proteins, tRNA synthetases as well as components of translation machinery, DNA repair and polymerases. The mRNAs corresponding to the PhyEco markers were extracted using the HMMer package (version 3.1.b1) (Eddy, 2011).

The mRNA design script (Umu et al., 2016), which is a wrapper for several other scripts, required five arguments: a specific gene of interest, a set of core mRNAs from species of interest which is used to calculate the codon adaptation index, a set of core ncRNA genes from the species of interest which is used for calculating avoidance, the entire set of all protein coding genes from the species of interest that is used to create the RNA codon distribution, and finally a numeric argument specifying the total number of mRNAs generated per possible avoidance region (e.g. if the numeric argument is 100 and the possible number of avoidance regions is 1000, approximately 100 x 1000 = 100,000 mRNA variants will be generated). All the datasets used to generate the variants mRNAs were in FASTA format. For our purposes the gene of interest was *araC* from *E. coli* strain REL606. We therefore obtained all additional mRNAs, core mRNAs and ncRNAs from this strain.

Determining Regions Significant for Avoidance

Previous research detected that the region of the transcript that was most significant for avoidance was the in first 21 nucleotides as estimated using a sliding window analysis (Umu et al., 2016). Sliding window analysis (SWA) is a commonly used method for investigating the properties of molecular sequences. The data that is plotted is a 'moving' average of a particular criterion based on window size and step size (Proutski & Holmes, 1998). The window size is a range that spans across an area of sequence and the step size determines how far along your sequence of interest your analysis will move to take each measurement. Umu et al, (2016) used a dataset of GFP mRNAs to determine regions along the GFP transcript with the greatest potential for interaction with ncRNAs. Their criterion of interest was therefore the affinity for interaction with ncRNAs among GFPs. The analysis calculated MFE values using a window size of 21 nucleotides and a step size of 1 nucleotide moving 5' to 3' along the GFP mRNAs. For each region, they computed the associated Spearman's correlation coefficients and *P* values. The analysis revealed the significance of first 21 nucleotides on protein expression which is consistent with previous findings that describe initiation as the rate limiting step in the translation of mRNA (Mao et al., 2014; Plotkin & Kudla, 2011; Tuller & Zur, 2015). Other regions of lesser significance along the GFP mRNAs were also revealed. A smaller scale test also detected regions of avoidance in the 5' UTR region proximal to the CDS. Based on these findings we chose to make alterations to our *araC* variants in the first 21 nucleotides. Additionally focusing on this small region of the transcript also eased computational complexity (Umu et al., 2016).

Determining Regions Significant for Secondary Structure

MFE values computed for mRNA folding predicted the stability of RNA secondary structures. The folding MFEs of variant *araC* mRNAs were calculated using the RNAfold algorithm (version 2.3.3) (Lorenz et al., 2011). The folding energy of our designed *araC* variant mRNAs was restricted to the first 37 nucleotides (Figure 2.3) as previous experiments with synonymously variant GFP mRNAs revealed that the most significant correlation between mRNA and protein was identified in this region (Kudla et al., 2009). Interestingly the folding energy of the entire GFP mRNA was not significantly correlated with GFP expression, however the folding energy of the first third of the mRNA was strongly correlated, indicating that mRNAs with stronger structure reduced expression of GPF.

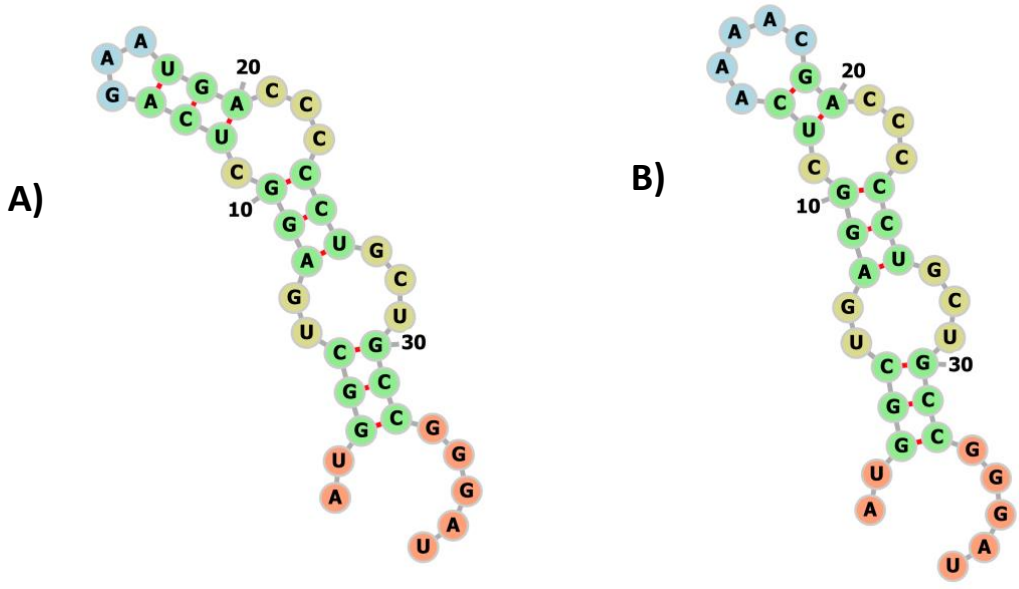


Figure 2.3: Predicted secondary structures within the first 37 nucleotides of synthetic and intermediate *araC* gene variants **A)** The intermediate *araC*INT start site, altered for an intermediate interaction with native ncRNAs in *E. coli*. **B)** The synthetic *araC*SYN start site, designed to have a high interaction with native ncRNAs in *E. coli*. These figures were produced using the RNAfold secondary structure drawing tool FORNA (Lorenz et al., 2011).

mRNA Design

The entire set of ncRNAs from *E. coli* B strain REL606 was run through a Python script developed by (Umu et al., 2016), along with our gene of interest, *araC*. The core ncRNAs used in this study consisted of 113 ncRNAs including rRNAs, tRNAs, 6S RNAs, RNaseP, SRP RNA and tmRNA. After filtering for redundant ncRNA sequences the summed avoidance scores were calculated based on a subset of 52 unique ncRNAs. To calculate the minimum binding energies of mRNA: ncRNA interactions we used RNAup (version 2.3.3) (Lorenz et al., 2011). The estimated level of ncRNA avoidance for each *araC* mRNA was computed by summing these binding MFEs. In other words, for each *araC* mRNA we computed 52 independent binding MFE values for each ncRNA and summed them to give a final avoidance MFE. A higher summed MFE score implying a higher avoidance and vice versa. Secondary structure was measured across the first 37 nucleotides and codon adaptation across the full length of the mRNA. In the same manner, a higher summed secondary structure MFE implies less intramolecular interaction and vice versa. CAI values ranged from 0 to 1, with higher values indicating a higher proportion of highly-adapted codons in the mRNA. In addition, G+C content was also strictly controlled for in gene design. Innate or intrinsic avoidance, as has been discussed by Umu et al, (2016) refers to an mRNAs innate features that restrict the mRNA from interacting with ncRNAs. This may pertain to the composition of nucleotides in the mRNA compared to ncRNAs. For instance, if the mRNA is composed mainly of G+C nucleotides, while ncRNAs are composed predominantly of A+U nucleotides, these two molecules will rarely interact.

We produced a massive sample of unique mRNAs that synonymously encode the *araC* gene. The *araC* mRNA variants were generated, from which we pooled a subset of 50 variants with the lowest summed MFE values for avoidance in the 5' region. From this pool, we selected 2 different mRNAs, a poor avoider, with the lowest overall minimum free energy value from the total population of generated variants and an intermediate avoider, that has a value that is midway between the native avoidance value of *araC* and the poor avoider. As we were only interested in testing avoidance we extracted the first 21 nucleotides from the low avoidance and intermediate avoidance variants. These 21 optimal nucleotides were used to replace the first 21 nucleotides in the native gene, leaving the region downstream of this start site unaltered This produced the two *araC* gene variants used in this study, *araC*SYN (low avoidance) and *araC*INT (intermediate avoidance) (Figure 2.4). Following this secondary structure and codon adaptation were re-calculated for these two variants.

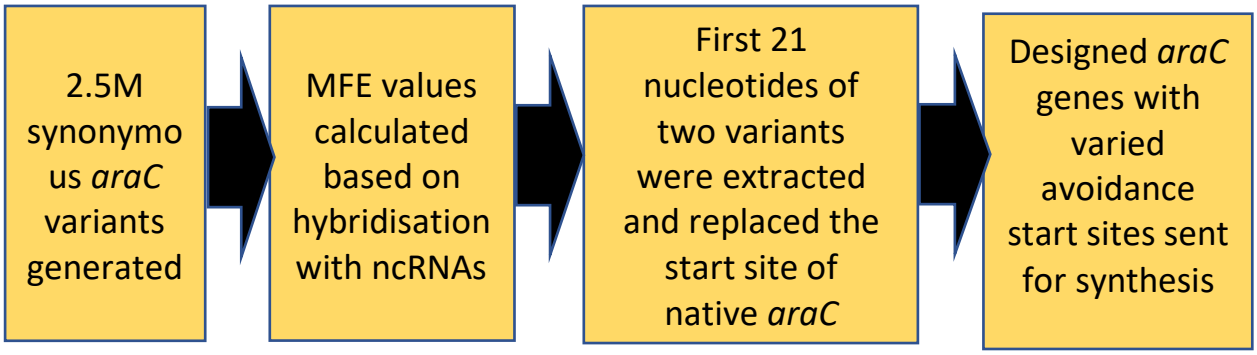
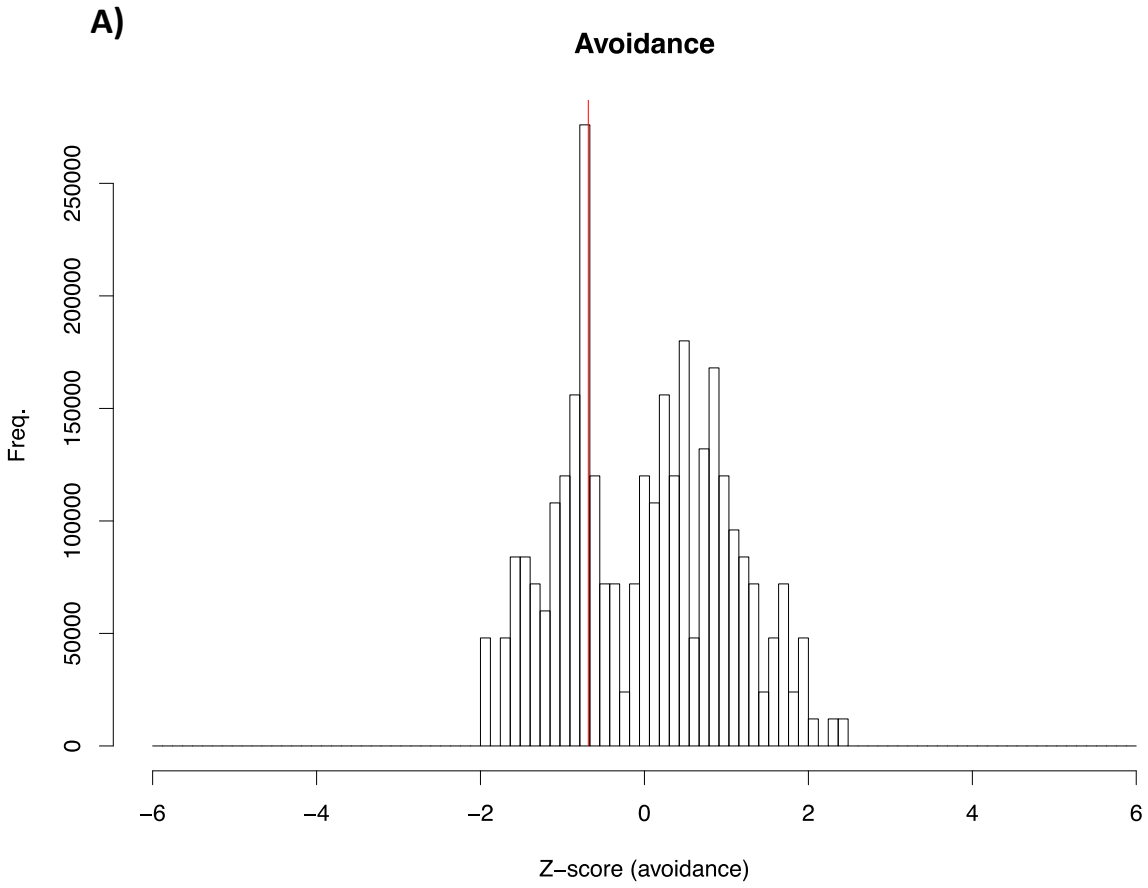
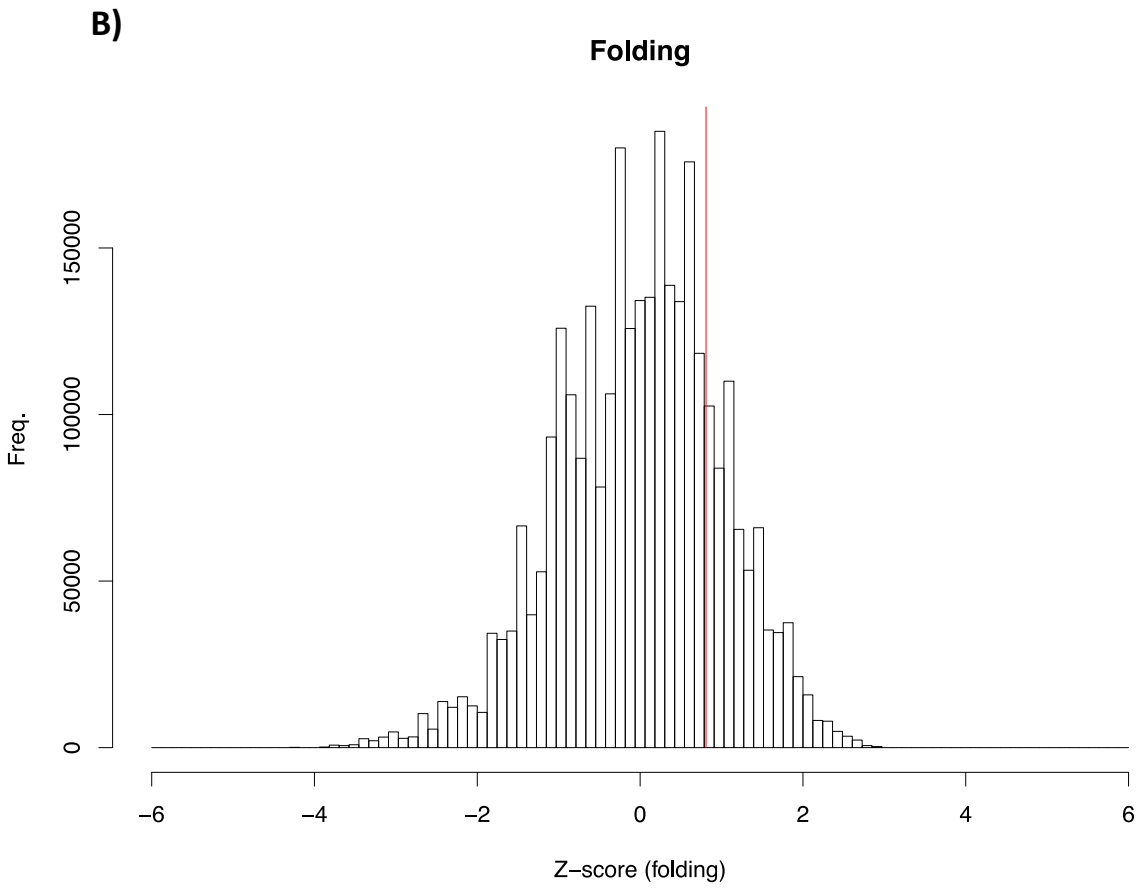


Figure 2.4. Workflow for the design of the variant *araC* genes used in our avoidance assay. The first 21 nucleotides of the *araC* gene variants designed for variation in avoidance MFE were extracted and replaced the start site of the native *araC* gene in the final constructs.

Results

We sampled 2.5 million synonymous mRNA variants of *araC* (this mRNA can be encoded by 1.39×10^{146} unique variants). The variants were generated based on both core protein coding genes and the entire array of protein coding genes within the *E. coli* genome. The generated mRNA variants were scored based on (1) codon adaptation across the full length of the mRNA, (2) secondary structure/internal folding of the 5' region and (3) avoidance of mRNA: ncRNA interaction in the first 21 nucleotides of the molecule which. For each design parameter of the mRNA variants Z-score distributions were computed (Figure 2.5). The entire set of protein coding genes from REL606 was used as our sampling distribution. Codon adaptation was measured using only the core genes, as this created a wider distributed CAI. The *araC* variants generated were designed to vary only in their potential for interaction with ncRNAs in *E. coli* while keeping codon adaptation and secondary structure near average.


A)
Avoidance
Freq.
0
50000
100000
150000
200000
250000
−6
−4
−2
0
2
4
6
Z−score (avoidance)


B)
Folding
Freq.
0
50000
100000
150000
−6
−4
−2
0
2
4
6
Z−score (folding)

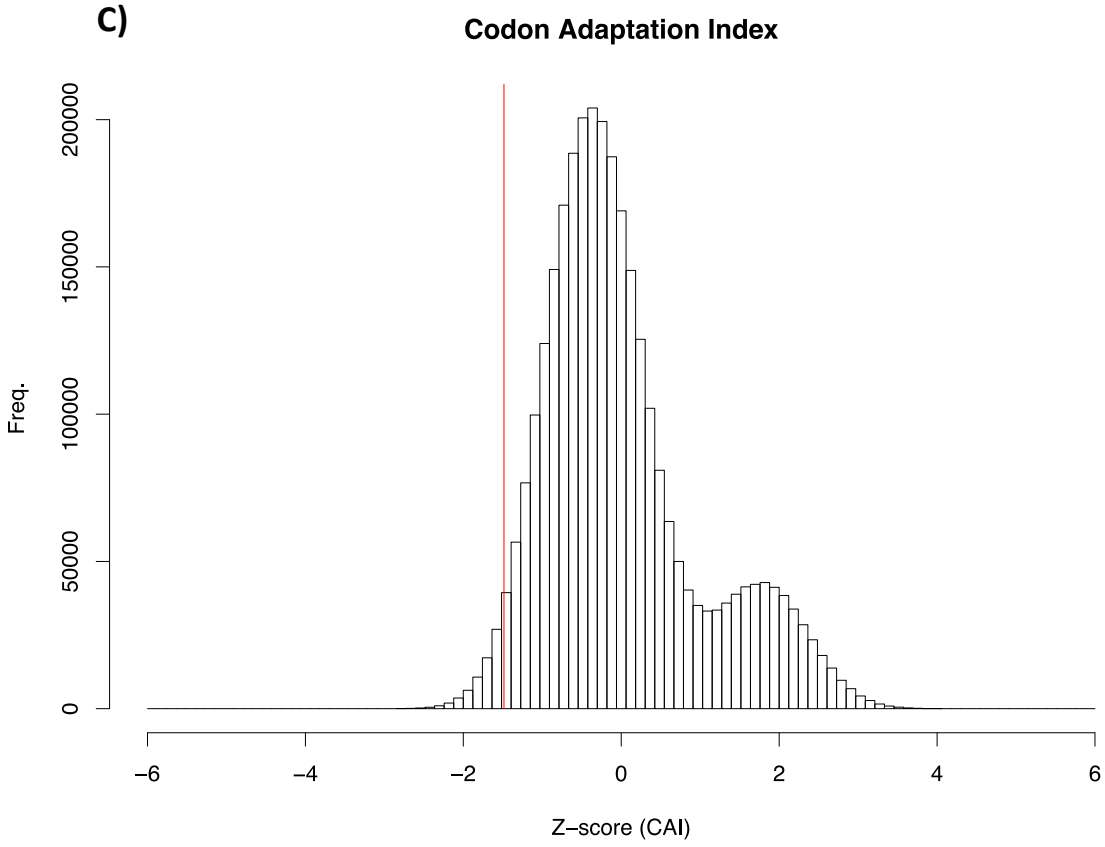


Figure 2.5. Z-score distributions for avoidance, secondary structure and codon adaptation index of designed *araC* mRNA variants. The red line indicates where the native *araC* mRNA sits in relation to the 2.5M generated variant mRNAs. Frequency indicates the number of mRNAs, Z-score indicates the number of mRNAs that fall within that Z-score range. **A)** Z-score distribution for avoidance for designed mRNA variants. Generated *araC* mRNAs fall above and below the native *araC* resulting in a bimodal distribution. **B)** Z-score distribution for secondary structure of designed mRNA variants. Most generated mRNAs fall around the mean (Z-score of 0). **C)** Z-score distribution for CAI of designed mRNA variants. Most generated mRNAs fall close to the mean (Z-score of 0).

The final variants were selected based on their overall avoidance MFE values in relation to the native *araC* mRNA in which the first 21 nucleotides of the mRNA produced an MFE value of -271.51 kcal/mol. Following calculation of this two specific *araC* variants were selected, an intermediate avoidance *araC* (MFE -306.7) and a low avoidance *araC* (MFE -342.04) (Table 2.1) relative to the native *araC*. As only the first 21 nucleotides of the two selected variants were implemented in the final design of the two *araC* gene variants, *araC*SYN and *araC*INT, secondary structure and CAI had to be re-calculated. MFE values for the secondary structure of the first 37 nucleotides were measured using the RNAfold algorithm (Lorenz et al., 2011). The native mRNA demonstrated the lowest potential for secondary structure (MFE -6.4) followed by the synthetic (low avoidance) mRNA (MFE -7.0) and the intermediate mRNA (MFE -7.4) (Table 2.1). CAI was also measured utilising the entire gene length for each mRNA variant. The proportion of adaptive codons for each variant and the wild type were very similar. The wild type *araC* was calculated to have a CAI of 0.579, the synthetic CAI was 0.5777 and the intermediate had a CAI of 0.5776. The final gene constructs, synthetic and intermediate, were extracted and sent to Macrogen to be synthesised for the scarless allelic replacement. Further analyses were performed to assess the potential level for interaction between native ncRNAs in *E. coli* and the structural genes of the arabinose operon, *araB*, *araA* and *araD*. The results of these analyses show that relative to *araC* the native potential for interaction in the structural genes of the arabinose operon is much lower than in *araC* itself (Table 2.2).

Table 2.1. The table shows the first 21 nucleotide sequence and the respective summed avoidance and secondary structure MFE scores for the native, intermediate and low avoidance *araC* gene variants. Secondary structure was calculated over the first 37 nucleotides.

Synonymous AraC Variant	Sequence (First 21 nts)	Avoidance MFE (kcal/mol)	2° Structure MFE (kcal/mol)
araC (Native)	ATGGCTGAAGCGCAAAATGAT	-271.5	-6.4
Intermediate (Med Affinity for ncRNAs)	ATGGCTGAGGCTCAGAATGAC	-306.7	-7.4
Synthetic (High Affinity for ncRNAs)	ATGGCTGAGGCTCAAAACGAC	-342.0	-7.0

Table 2.2. The table shows the structural genes of the arabinose operon, *araB*, *araA*, *araD* and *araC*. For each gene, the avoidance MFE, secondary structure MFE and CAI metric have been calculated.

Arabinose Operon Gene	Avoidance MFE of First 21 nts (kcal/mol)	Secondary Structure MFE of First 37 nts (kcal/mol)	Codon Adaptation Index CAI of the Entire Gene Length
araB	-193.6	-5.4	0.69
araA	-168.6	-3.0	0.76
araD	-153.2	-3.4	0.68
araC	-271.5	-6.4	0.58

Chapter 3
Generating *araC* Variant Strains of REL607

The design for the synthetic, *araCSYN* (low avoidance) and intermediate, *araCINT* (moderate avoidance) constructs were sent to Genscript (Piscataway, New Jersey) to be synthesized. The synthesised constructs we received back were integrated into the standard molecular cloning plasmid pUC57. Each construct was designed to include 500bp homology arms to aid recombination in later steps. The original protocol was modified slightly so that sites flank the homology arms for ease of integration of the constructs into the temperature sensitive plasmid pST76-A in the latter steps of this protocol. The pUC57::*araC*SYN (low avoidance) and pUC57::*araC*INT (intermediate avoidance) plasmids were transformed into the DH5α strain of *E. coli.* The constructs were amplified by PCR with primers carrying restriction sites that flanked each end of the homology arms, +500bp upstream and -500bp downstream of the initial 21 nucleotides of the *araC* gene. A scarless allelic replacement of the *araC* gene was then preformed (Fehér et al., 2008). The synthetic and intermediate constructs were then cloned into the multiple cloning site (MCS) of the pST76-A plasmid (Pósfai, Koob, Kirkpatrick, & Blattner, 1997). The constructs were cut with restriction enzymes, EcoRI (Invitrogen) and XmaI (Invitrogen), corresponding to the restriction sites incorporated during the PCR step. Digested constructs were added to a ligation reaction containing the pST76-A plasmid, which was cut with the same restriction enzymes, to incorporate them into the MCS of temperature sensitive plasmid. The ligated products were transformed into DH5α using bacterial transformation. The transformation process saw the cells were treated with calcium chloride ($CaCl_2$) and heat shocked at 42°C to enable the plasmids to enter the cell. The cells were then grown at 30°C which is permissible for replication of the plasmid. The temperature sensitive

nature of pST76-A was exploited by growing the cells at temperatures non-permissible for replication of the plasmid (42°C). Colonies were streaked onto media LB+amp+sm and incubated at 42°C forcing the cells to integrate the plasmid carrying the gene constructs into the chromosome of REL607 (Lenski, 1988). pST76-A carries an ampicillin (amp^r) resistance marker thus only those cells that integrate the plasmid into the chromosome could grow on the media. Following the integration of pST76-A into the chromosome REL607 was transformed with a second helper plasmid, pSTKST, which carried I-SceI, a restriction enzyme that is induced by the introduction of chlortetracycline. I-SceI induces recombination by cutting the DNA at a recognition site along the pST76-A plasmid. RecA-mediated recombination resulted in a 50:50 ratio of either the original wild-type sequence or the replacement gene. Subsequent growth experiments were performed to determine the effect of the gene replacement on growth of *E. coli* under selective pressure to express *araC*.

Strains and Media

Several different media were used across this study. *Escherichia coli* B strain REL606 was obtained from Tim Cooper (University of Houston, Texas) and REL607 was a spontaneous revertant of REL606 generated in the lab. For the knock-in protocol Luria Bertani (LB) media was used. For solid media LB agar was prepared using bacteriological agar that was added to a concentration of 1.5% w/v (HyAgarose). Tetrazolium arabinose agar was prepared in distilled water using the following ingredients; tryptone 1% w/v, yeast extract 0.1% w/v, sodium chloride 0.5% w/v and bacteriological agar 1.6% w/v. The following antibiotics were also added to media where necessary at the following concentrations: streptomycin, 100µg/ml; ampicillin, 100µg/mL; kanamycin, 20µg/mL.

Description of plasmids

The following plasmids were used as part of the scarless allelic replacement. A brief description of each plasmid follows:

- pST76-A is a temperature-sensitive plasmid replicon as such it cannot replicate at 37–42°C. In addition to an ampicillin resistance marker this plasmid also carries an I-SceI restriction site that is cut by the I-SceI restriction enzyme (Figure 3.1 A).
- pSTKST is a helper plasmid that when induced by chlortetracycline expresses I-SceI, a restriction enzyme. This plasmid also carries a kanamycin resistance marker (Figure. 3.1 B)

A) pST76-A **B) pSTKST**

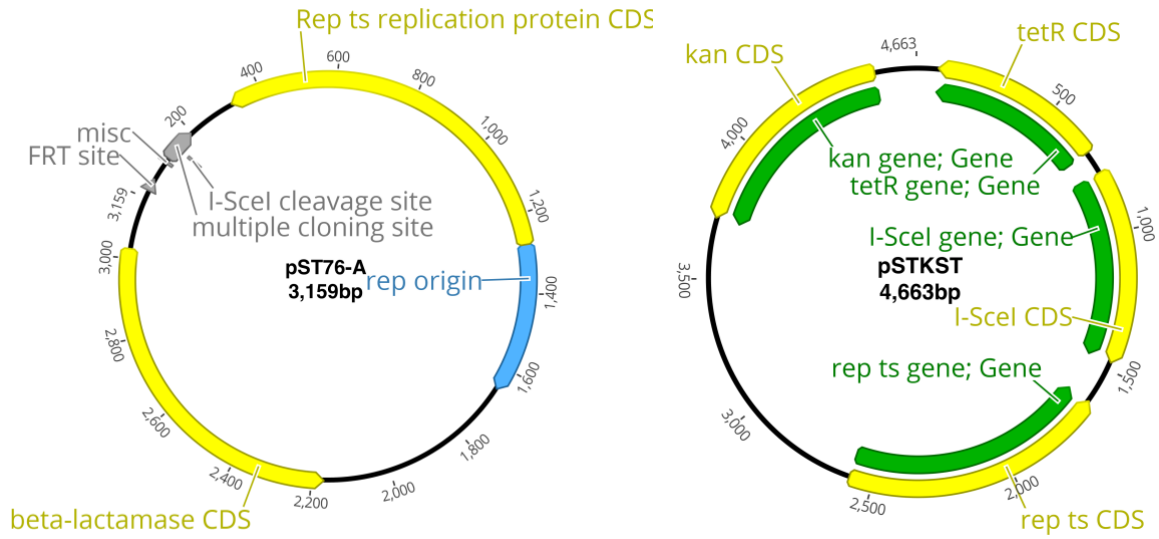


Figure 3.1. Plasmid maps of pST76-A and pSTKST, the two key plasmids in the scarless allelic replacement protocol. **A)** The *araC* gene varaints were cloned into pST76-A at the MCS to later be integrated into the chromosome of REL607 **B)** pSTKST was used to induce recombination via expression of I-SceI which cleaves pST76-A allowing recombination to occur.

Transformation of DH5α Strains with pUC57:*araC* Variant Plasmids

To insert the DNA of *araC* gene variants into DH5α cells I performed a bacterial transformation as described by (Sambrook, Russell, & Maniatis, 2001). A day culture of DH5α (Table 3.1) was created by diluting DH5α overnight culture 1:100 into 100ml of LB and growing the cells at 37°C to OD ~0.6 for 2-4 hours. These cultures were then cooled on ice for 10mins. Cells were spun down at 6000rpm for 10 minutes at 4°C to form a pellet. The cell pellet was then resuspended in 30ml of ice cold $CaCl_2$. 100µl of cells were aliquoted into each of 4 Eppendorf tubes. Into each tube was added 1µl of either (1) synthetic pUC57 plasmid (2) intermediate pUC57 plasmid (3) control plasmid pBR322 or (4) no plasmid. The cells were then incubated on ice for 30 minutes, heat shocked at 42 °C for 90 seconds followed by another incubation step on ice for 5 minutes. 10% and 90% dilutions of each tube were plated onto LB+amp plates and incubated overnight at 30°C.

Table 3.1. Strains used in this study. The name of the strain is presented on the left with the corresponding genotype on the right.

Strains	Genotype
DH5α	F– Φ80*lac*ZΔM15 Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17 (rK–, mK+) *pho*A *sup*E44 λ– *thi*-1 *gyr*A96 *rel*A1
REL606	F-, tsx-467(Am), araA230, lon-, rpsL227(strR), hsdR-, [mal+](LamS)
REL607	F, tsx-467(Am), lon- , rpsL227(strR), hsdR- , [mal+](LamS)
REL607::araC_SYN	F-, tsx-467(Am), lon-, rpsL227(strR), hsdR- , [mal+](LamS), insertion of low avoidance *araC* start site
REL607::araC_INT	F-, tsx-467(Am), lon-, rpsL227(strR), hsdR- , [mal+](LamS), insertion of intermediate avoidance *araC* start site

PCR Screening of DH5α Colonies Transformed with pUC57::*araC*

Screening of transformed DH5α was performed using PCR. 1:10 dilutions of M13 forward and reverse primer stocks (**Table 3.2**) in molecular grade water were made to a total final volume of 100µl. A master mix containing KAPA2G Robust Hot Start Ready Mix (KAPA Biosystems), molecular grade water, and the M13 forward and reverse primer dilution was prepared for the PCR reaction. KAPA2G Hot Start Ready Mix is a ready-made PCR mixture containing: taq polymerase, dNTPs, reaction buffer and $MgCl_2$. 10µl of the master mix was aliquoted into PCR tubes. Single colonies were transferred via grid plating into PCR tubes using a pipette tip.

PCR was performed under optimal conditions adjusting the cycling program for the primers used. The PCR protocol used was as follows: Initial denaturation 95°C, 3 minutes. Then 30x cycling: Denaturation: 95°C, 15 seconds, Annealing: 58°C, 15 seconds and Extension: 72°C, 1minute 30 seconds. Then: Final extension: 72°C, 5 minutes

PCR products were verified using Gel Electrophoresis. The gels were a 1% agarose solution (HyAgarose) dissolved in 1X Tris Borate EDTA (TBE) Buffer with 7µl/ml of SYBR Safe (Invitrogen) for staining of DNA fragments. The Generuler 1kb DNA ladder (Invitrogen) was prepared and run in parallel with the samples. The gels were maintained in 1X TBE buffer during electrophoresis at 110V for 20-25 minutes. Fragment sizes were visualised by exposure to UV light using the BIO RAD Molecular Imager Gel Doc XR System with Image Lab software (version 5.2.1).

PCR Amplification of *araC* Construct DNA

PCR of the *araC* Construct DNA was performed using the high-fidelity polymerase KAPA HiFi (KAPA Biosystems). A master mix containing KAPA HiFi, molecular grade water, EcoRI_*araC*-500bp and XmaI_*araC*+500bp (Table 3.2) was made. The *araC* -500bp and + 500bp primers (Table 3.2) carry EcoRI and XmaI restriction sites at their 5' ends respectively. Restriction sites were incorporated into the design of the *araC* primers to facilitate the restriction enzyme digest and re-ligation into pST76-A in later steps of the knock-in protocol. 50µl of the master mix was aliquoted into PCR tubes followed by 1µl of the *araC* constructs.

Table 3.2. Primers and other oligonucleotides. Primers used for obtaining *araC* avoidance-variant knock-ins. The name of the primer is presented on the left with corresponding primer sequence on the right.

Primer	Sequence
araC KO-d	5'-GAATAAATACCGCCAATATAGC-3'
araC KO-e	5'-GCAAAATATCGATATACACCGGC-3'
araC -64 bp	5'-ACGGCAATGTCTGATGCAATAT-3'
araC +142 bp	5'-CGCCATCAATGAATACACGGTAG-3'
araC +37 bp	5'-ACTCCGTCAAGCCGTCAATT-3'
araC -721 bp	5'-GCTTCTTCAACCGCAGTGTG-3'
EcoR1-araC -500bp	5'-GGACTT GAATTC TGCCGCTTCCATTGACTCAA-3'
XmaI-araC +500bp	5'-GGACTT GGGCCC CCTCGCGTACCCGATTATCC-3'
pSTKST F	5'-GGACCT GAATTC TTTCCCCAAAAGTGCCACC -3'
pSTKST R	5'-GGACCT GAGCTC GACGAGTTCTTCTGAGCGGG-3'
T1	5'-CGGAAGGATCTGAGGTTCTTATGGC-3'
T2	5'-CGAATTGTCGACAAGCTTGATCTGGC-3'
M13 F	5'-GTAAAACGACGGCCAGT-3'
M13 R	5'-AGCGGATAACAATTTCACACAGGA-3'

Preparation for Scarless Knock-In

A suicide plasmid based gene-replacement method that utilises homologous recombination (Fehér et al., 2008) was used to replace the wild type *araC* gene of REL607 with our engineered *araC* avoidance variants, *araC*SYN *and araC*INT. The 1 kb-long DNA constructs carrying the altered *araC* start sites were amplified using KAPA Hi-fi Polymerase with primers flanking either side of the 21-nucleotide region (500 bp upstream and downstream). The primers were designed to carry restriction site at their 5' ends to aid the efficiency of cloning in later steps. Purification of PCR products containing the *araC* variants constructs *araC*SYN and *araC*INT (synthetic and intermediate respectively) was carried out using the Wizard SV Gel and PCR Clean-Up System (Promega). Purification of the PCR product included the removal of taq polymerase, primers, dNTPs and buffer solution. DNA was quantified using the NanoPhotometer® (IMPLEN).

Restriction Endonuclease Digestion and Ligation of Constructs into pST76-A

The 1kb fragments carrying the altered *araC* start sites and the temperature sensitive plasmid pST76-A were digested with restriction endonucleases EcoRI (Invitrogen) and XmaI (Invitrogen). Two separate digests were set up for this step (1) the low avoidance PCR product and (2) the intermediate avoidance PCR product (Table 3.3) Reactions were set up in PCR tubes and placed at 37°C for 16 hours.

Ligation of the *araC* gene constructs into the pST76-A plasmid vector was performed using the Rapid DNA Ligation Kit (Invitrogen) containing a T4 DNA ligase with 5x Rapid DNA Ligation

Buffer. This protocol used an insert: vector ratio of 3:1 to a total reaction volume of 20µl. The ligation was carried out with several controls in place to ensure proper ligation takes place. Controls are as follows:

1. 16µl water, 4µl buffer (growth indicates contamination)
2. 2µl undigested pST76-A vector, 14µl water, 4µl buffer (ensures the plasmid works)
3. 2µl restriction digested pST76-A vector, 14µl water, 4µl buffer (expect no growth unless the vector has been cut improperly)
4. 2µl restriction digested pST76-A vector, 13µl water, 4µl buffer, 1µl T4 DNA ligase (growth here indicates the vector has re-ligated on itself)

The intended sample contained 2µl restriction digested pST76-A vector, 7µl water, 4µl buffer, 1µl T4 DNA ligase and 6µl of *araC* construct insert. The ligation mixture was left at room temperature overnight to allow sufficient time for ligation to occur.

Table 3.3. Restriction Enzymes Digest of *araC* constructs. Two separate digests were set up. Each digest was carried out in the same manner. Use of XmaI buffer meant that EcoRI only had 50% activity and was therefore added in a double dosage.

Digested DNA	Digestion Mix	Enzymatic Activity in XmaI Buffer
***araC*SYN Construct**	pST76-A, 2x EcoRI, XmaI, XmaI Buffer, Molecular Grade Water	EcoRI 50% XmaI 100%
***araC*INT Construct**	pST76-A, 2x EcoRI, XmaI, XmaI Buffer, Molecular Grade Water	EcoRI 50% XmaI 100%

Transformation of DH5α Strains with pST76-A::*araC* Variant Plasmids

Competent DH5α cells were thawed on ice and then had 3µl of ligation mixture was then added. The cells were then incubated on ice for 30 minutes and then heat shocked at 42°C for 90 seconds (Sambrook et al., 2001). Cells were then cooled on ice for 5 minutes after which 900µl of LB was added. The cells were then left to recover for 1 hour. Tubes containing cells in each of the ligation mixtures were plated on LB+amp plates and incubated overnight at 30°C. Incubated plates were checked for growth and colonies were PCR amplified via grid plating. Single colonies were stabbed onto a grid plate before being transferred to PCR tubes containing a PCR master mix. The master mix consisted of KAPA2G, molecular grade water, as well as T1 and T2 primers (Table 3.2). These primers span the MCS of pST76-A. Product sizes were confirmed using gel electrophoresis, as previously described (see this Chapter, pg. 44).

Integration of pST76-A::*araC* Variant Plasmids into REL607 Chromosome

pST76-A integrated with the synthetic (*araC*SYN) and intermediate (*araC*INT) *araC* constructs were extracted from overnight cultures of DH5α + pST76-A::*araC* variants using the PureLink® Quick Plasmid Miniprep Kit (Invitrogen). DNA was quantified using the NanoPhotometer® (IMPLEN). pST76-A::*araC* variant plasmids were transformed into separate lines of REL607 with calcium chloride and heat shock (Sambrook et al., 2001). REL607 cells were cultured in LB+amp+sm and grown at 30°C for 4 hours, followed by growth at 42°C for 12 hours to integrate the plasmids into the chromosome. Following the integration step cell cultures were

then transferred to 37°C for 24 hours. Confirmation of the pST76-A::*araC* variant plasmids integration into the chromosome of REL607 was performed using a combination of primers. Theoretically the plasmid can integrate in two different orientations, at either end of the 500bp homology arms where there is homology with the *E. coli* chromosome. The primer combinations used were (1) araC -721bp and T1 (1.4kb), and (2) araC +37bp and T2 (1.5kb). As the products of each respective PCR are similar in size they were difficult to distinguish in a gel. The integration of the plasmid carrying constructs into the chromosome was therefore confirmed by Sanger sequencing (Macrogen).

Inducing Scarless Allelic Replacement

REL607 + pST76-A::*araC* was transformed with the helper plasmid pSTKST with calcium chloride and heat shock (Sambrook et al., 2001). Transformation was carried out as previously described and transformants were confirmed by Sanger sequencing (Macrogen). The knock-in was induced by growing up the REL607::pST76-A::*araC* + pSTKST variant strains in the presence of chlortetracycline (CTc) at 30°C for 24 hours in the following media: 10ml LB, 10µl Kanamycin, 20µl Streptomycin and 250µl Chlortetracycline.

10^0 and 10^2 dilutions were plated on plates containing the same media and were again grown overnight at 30°C. Theoretically half of the resulting colonies confer the desired construct (REL607::araC_SYN or REL607::araC_INT) + pSTKST with the other half conferring the WT sequence. Twenty positively screened colonies were subsequently confirmed by Sanger sequencing (Macrogen). pSTKST was later removed by growing the strains overnight at 42°C.

Growth Experiments

By growing the engineered *E. coli* strains on media which necessitates strong expression of the *araC* gene, (i.e. media supplemented only with minimal arabinose) we created a selective pressure for our engineered lines to adapt, creating a selective pressure to alter their *araC* gene sequences such that ncRNA interactions are minimised, increasing expression and enabling them to better utilise arabinose as a carbon source.

To measure the optimal arabinose concentrations for assaying growth differences between strains we used 12-well plates. Into each well 1ml of Davis-Minimal Media supplemented with differing concentrations of arabinose (0.02%, 0.2%, 1% and 2%) was added. Each plate was divided into four sections for measuring growth of 4 different *E. coli* strains (Figure 3.2), REL606, REL607, REL607::*araC*SYN and REL607::*araC*INT (Table 1.1). For this assay REL606 and REL607 represent negative and positive controls respectively. REL606 is an *Ara*$^{-}$ mutant making it unable to metabolise arabinose. REL607 carries the unaltered wild type *araC*.

1:100 dilutions of liquid culture were transferred into the wells containing 1ml of fresh Davis minimal media supplemented with arabinose. Each strain was cultured at each concentration. The OD595 of these cultures was monitored over 24-hours at 37°C (with shaking at 200rpm), and measurements were taken every 6 minutes using the FLUOstar OMEGA plate reader (BMG Labtech). Multiple replicates were performed at different concentrations. OD measurements and statistics for each strain were analysed and generate using the GrowthRates package (version 3.0) (Hall, Acar, Nandipati, & Barlow, 2014).

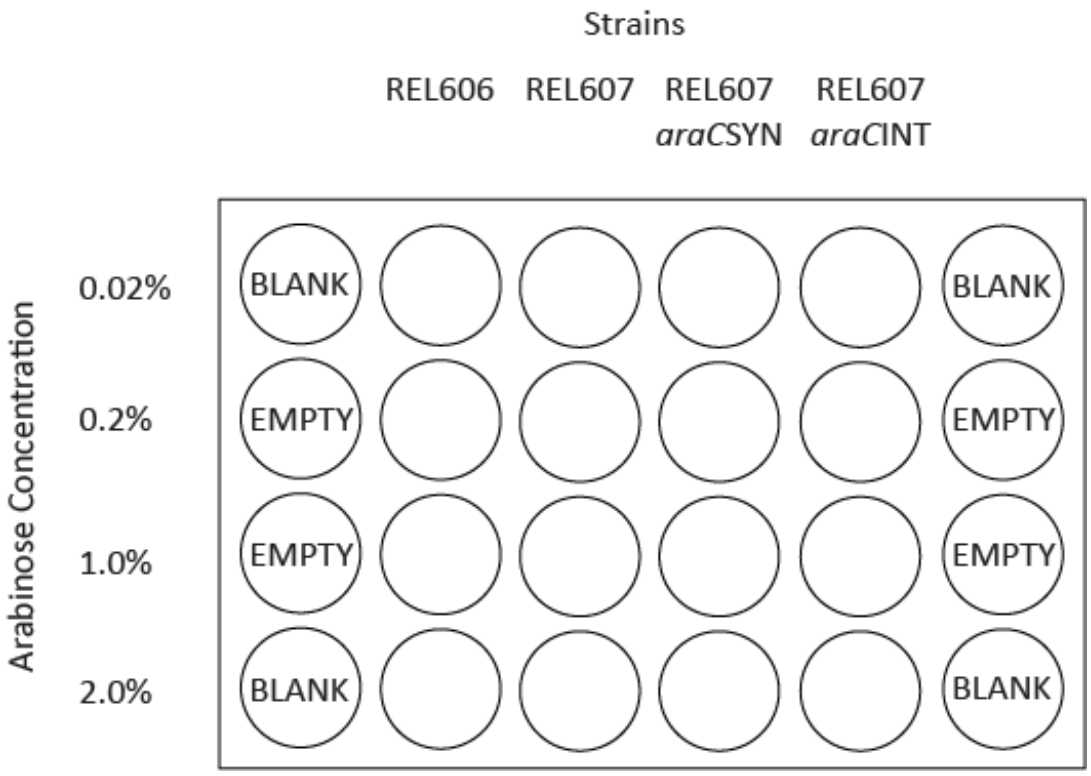


Figure 3.2. 24-well plate used for comparing growth between variant strains. The growth of the two variant strains, *araC*SYN and *araC*INT were compared. REL606 was used a negative control and REL607 as a positive control. Four different concentrations of arabinose were used for the initial growth assay, 0.02%, 0.2%, 1% and 2%. BLANK indicates a blanked well that contain only media. Blank wells are used to estimate the background absorbance of media. EMPTY indicates wells that contain no media and no culture.

Results

Scarless Allelic Replacement Preparation

To confirm that our gene constructs would be viable in *E. coli* DH5α was transformed with the pUC57 plasmids carrying the *araC* variants *araC*SYN and *araC*INT (low avoidance and intermediate avoidance) that we received from GenScript. Successful transformants were selected by plating the cells onto LB + ampicillin plates (pUC57 is ampicillin resistant) and

were incubated overnight at 37°C. Any colonies that formed were grid plated and screened using PCR and gel electrophoresis. M13 primers were used to screen the variant *araC* inserts cloned into pUC57 and gel electrophoresis revealed the expected product size of 1225bp (Figure 3.3) Grid plating allowed us to refer to the plate and grow up cultures of the colonies that produced the strongest bands in the gel, as these colonies provided optimal amplification of the PCR product.

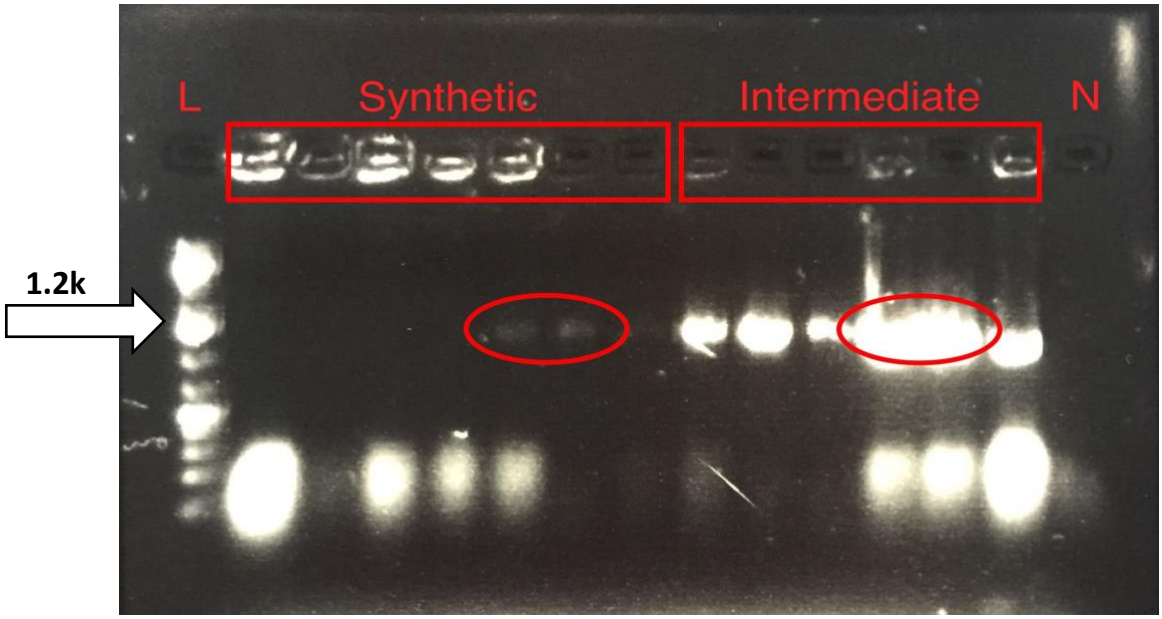
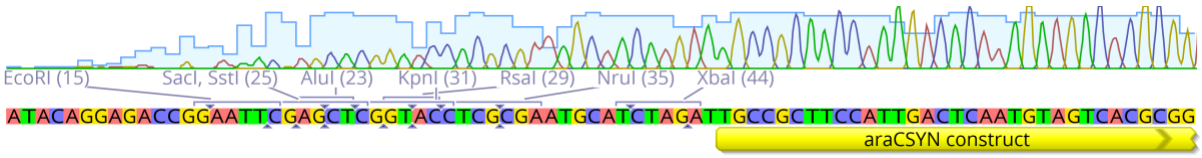


Figure 3.3. Gel image of PCR reaction of DH5α cells containing the araC variants, synthetic on the left, intermediate on the right. The ellipses indicate the two strongest bands from each *araC* variant. L indicates the 1kb GeneRuler ladder (ThermoFisher) that was used. N indicates the negative control which contained just the PCR master mix but no DNA.

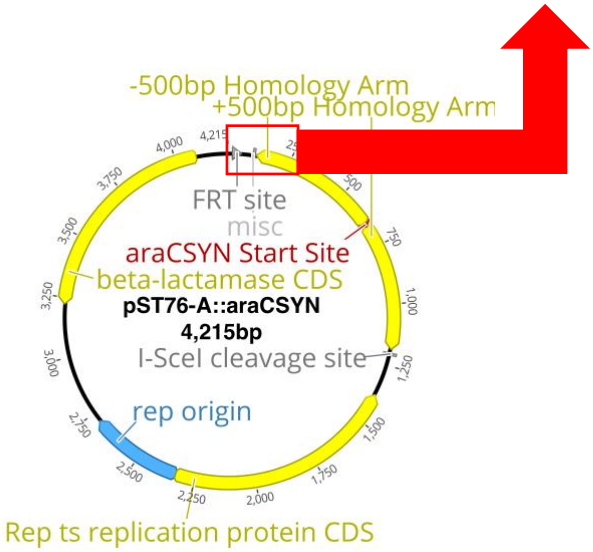
Digestion and Ligation of araC Variants and pST76-A Produced pST76-A::araC Variant Plasmids

The pUC57 plasmid harbouring the *araC* gene constructs, now transformed into DH5α, were then isolated so the constructs could be amplified for introduction into the suicide plasmid pST76-A. The constructs were amplified with PCR using primers flanked with restriction sites, and ligated into the MCS of pST76-A. Direct PCR of the ligation mixture revealed band sizes corresponding to the size of the MCS of pST76-A incorporated with the *araC* constructs (1.35kb). Incorporation of the *araC* constructs into the plasmid was confirmed with Sanger sequencing (Macrogen) (Figure 3.4). Two sequencing runs were performed per construct for confirmation. The first was run using the T2 forward primer and the second using the T1 reverse primer. The pST76-A::*araC* variants were then used to transform REL607. Successful transformants were subsequently grown at temperatures non-permissible for replication of the plasmid. This forces REL607 to integrate the plasmid into the chromosome where it can be replicated, thus only cells successful in integrating the plasmid will survive. Colony growth indicated that pST76-A had been successfully integrated into the chromosome. This was also verified using PCR. Gel images revealed consistent band sizes across multiple cultured replicates.

A) Sequencing Results

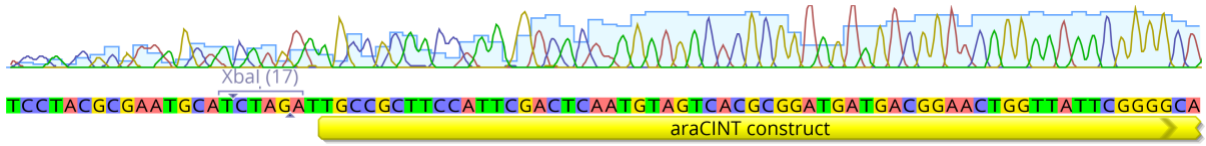


B) pST76-A::*araC*SYN

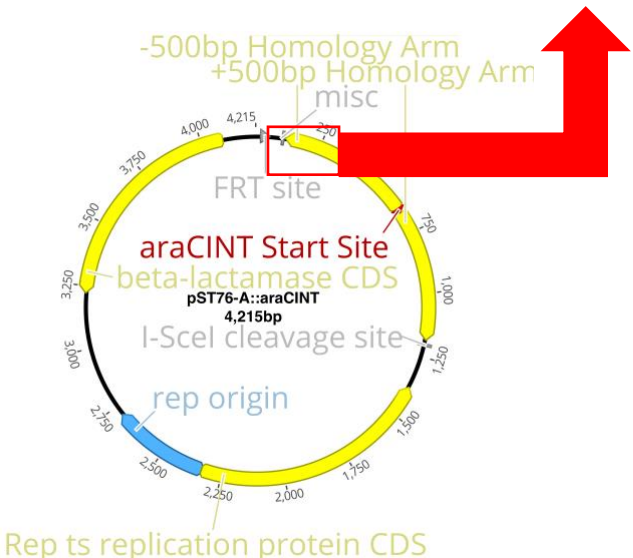


region shown

C) Sequencing Results



D) pST76-A::*araC*INT



region shown

Figure 3.4: Sequencing of pST76-A shows the correct cloning of the araC constructs into pST76-A creating pST76-A::*araC*SYN and pST76-A::*araC*INT. **A)** Sequencing results of pST76-A using T1 and T2 primers (Table 3.2) revealed the correct insertion of the *araC*SYN into the MCS of the plasmid. **B)** Cloning of *araC*SYN construct into pST76-A generated the pST76-A::*araC*SYN plasmid, which would later be transformed into REL607. **C)** Sequencing results of pST76-A revealed the correct insertion of *araC*INT into the MCS of the plasmid. **D)** Cloning of *araC*INT construct into pST76-A generated the pST76-A::*araC*INT plasmid.

Primer Medley Results

Following the induction step of the scarless protocol it was necessary to sequence the regions either side of the recombination site to ensure the *araC* gene variants had been integrated correctly. This required amplification of the *araC* gene. However, this proved rather difficult. We attempted to use primers that would amplify the region -721bp downstream and +37bp upstream of the 21-nucleotide gene start site but upon visualising the gels following gel electrophoresis either no bands could be detected or the bands were unspecific (Figure 3.5). Multiple PCRs were run using these primers however the result was the same. We then used different combinations of primers that amplify *araC* at different regions surrounding the gene. This method allowed us to identify primer combinations that produced a positive result (Figure 3.5). The final primers that were chosen to amplify this region for sequencing were the araC -721bp and +142bp (Table 3.2) as they could produce a clear band in the gel as well as provide sequence information on the integration of the *araC* constructs.

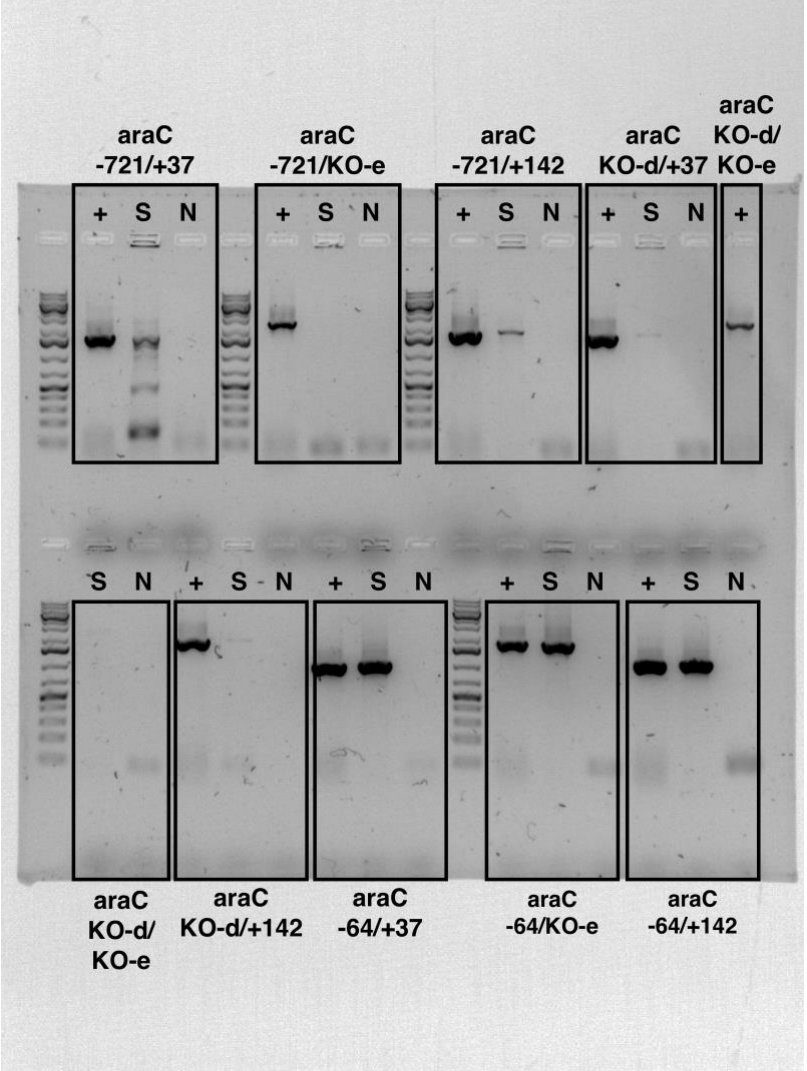


Figure 3.5. Gel image of the results the primer medley. All *araC* primers were used in this assay. **S** indicates the sample that was tested, **+** and **N** indicate the positive and negative controls respectively. The original primer pair, araC -721/+37, show unspecific bands in the gel. The final combination of primers used to amplify the *araC* gene region were araC -721/+142.

Inducing I-SceI Results in Replacement of araC in REL607::pST76-A::araC Variant Lines

Cells successful in integrating the pST76-A::*araC* variant plasmids into the chromosome of REL607 were then transformed with a second helper plasmid, pSTKST. pSTKST was used for the induction of the I-SceI gene which promotes recombination in the region where pST76-A is integrated by cleaving the I-SceI recognition site on the plasmid. RecA-mediated recombination then resulted in a 50:50 ratio of either the wild type araC or the replacement genes araCSYN and araCINT. Three sequencing runs were performed. The first was run using the araC -721bp forward primer to confirm correct integration of the araC gene construct into the intended site (Figure 3.6). The second was run using the araC -142bp forward primer. Finally, the third sequencing run was run in the reverse direction using the araC +64bp primer. This ensured we had sufficient sequence data to confirm the allelic replacement.

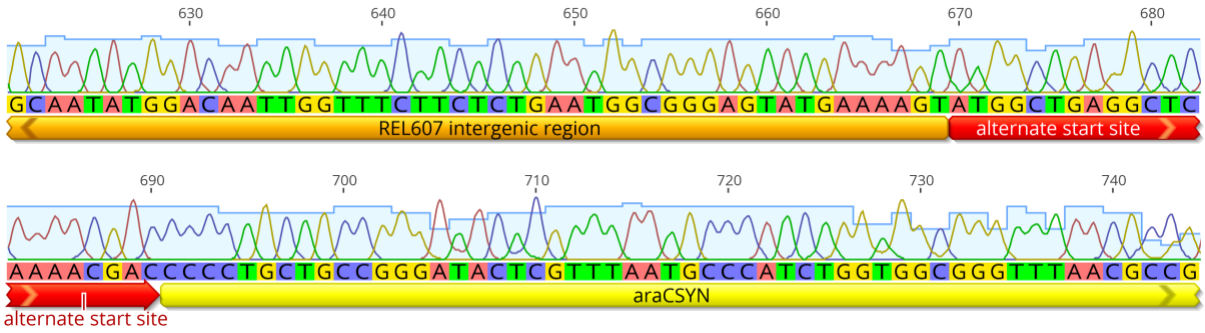


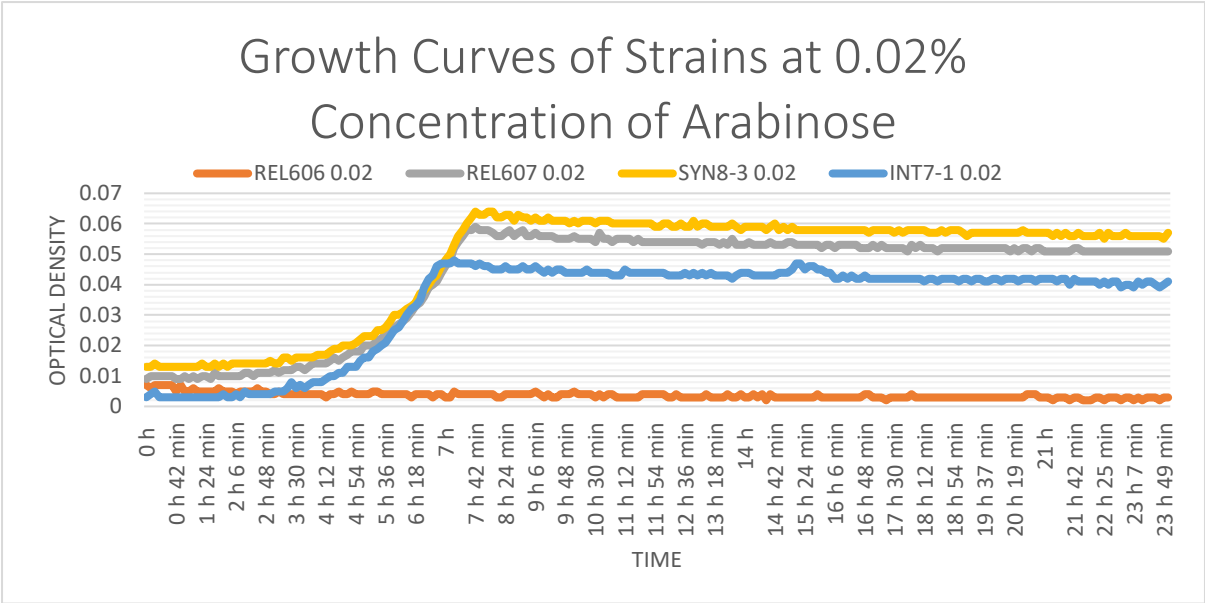
Figure 3.6. Sequencing results of *araC*SYN and *araC*INT in REL607 following the recombination step revealed that the variant *araC* gene constructs had been integrated correctly. The chromatogram was produced by Sanger sequencing of the *araC*SYN gene using the araC -721 forward primer.

Growth Rates of REL607-SYN/INT Lines are Indistinguishable from Wildtype REL607

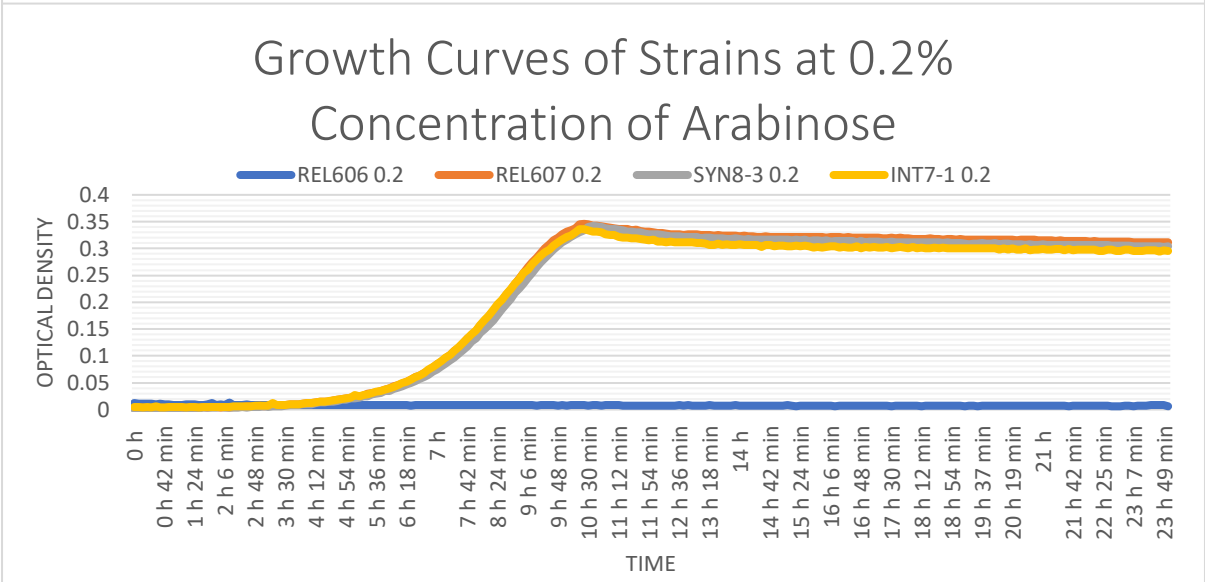
Growth rates of our bacterial strains were determined based on optical densities measured over a 24-hour growth period. First examination of the growth curves indicated that there were minimal differences in the rate of growth between strains (except for the REL606) when grown at varying concentrations of arabinose (Figure 3.7). This was confirmed upon performing a Student's t-test which revealed no significant difference in the doubling times between REL607 and the *araC* variant strains REL607::*araC*SYN (t-test, p = 0.548) and REL607::*araC*INT (t-test, p = 0.0974) respectively at a 2% concentration of arabinose. Additionally, the same test revealed no significant difference in the growth rates between the two engineered strains REL607::*araC*SYN and REL607::*araC*INT (t-test, p = 0.295). As expected the negative control strain, REL606, saw a much lower absorbance level than the either of the designed strain or the wild-type. To eliminate the possibility that in the generated strains being tested some mutation may have arisen that might affect their ability to interact with ncRNAs additional growth assays were performed to determine if there was any significant difference in growth rates between replicates of the generated strains. To test this four replicates of the *araC*INT strain (INT-1, INT-3, INT-4 and INT-5) and four replicates of the araCSYN strain (SYN8-3, SYN8-4, SYN8-7 and SYN8-8) were grown in media containing a 0.02%, 0.2% or 1% concentration of arabinose. Analysis of the growth curve data implied there was no difference in growth (Figure 3.8 A-F). Further growth assays were performed with multiple replicates of SYN 8-3 and INT 7-1 in the same 24-well plate (Figure 3.8 G) again the growth curves indicated minimal differences between the strains. A single factor analysis of variance (ANOVA) confirmed this revealing no significant difference between the growth rate of REL607-INT strains under these conditions (p = 0.0553) although, this p-value indicates

some level of variation. To verify whether this variation was significant a second ANOVA was performed, this time using doubling times calculated from the GrowthRates package however this was not significant ($p = 0.123$) Additionally, a single factor ANOVA of the REL607-SYN strains also revealed no significant differences between the growth rates of these replicates ($p = 0.391$). ANOVA was also performed using doubling time which also revealed no significant differences between replicates ($p = 0.696$)

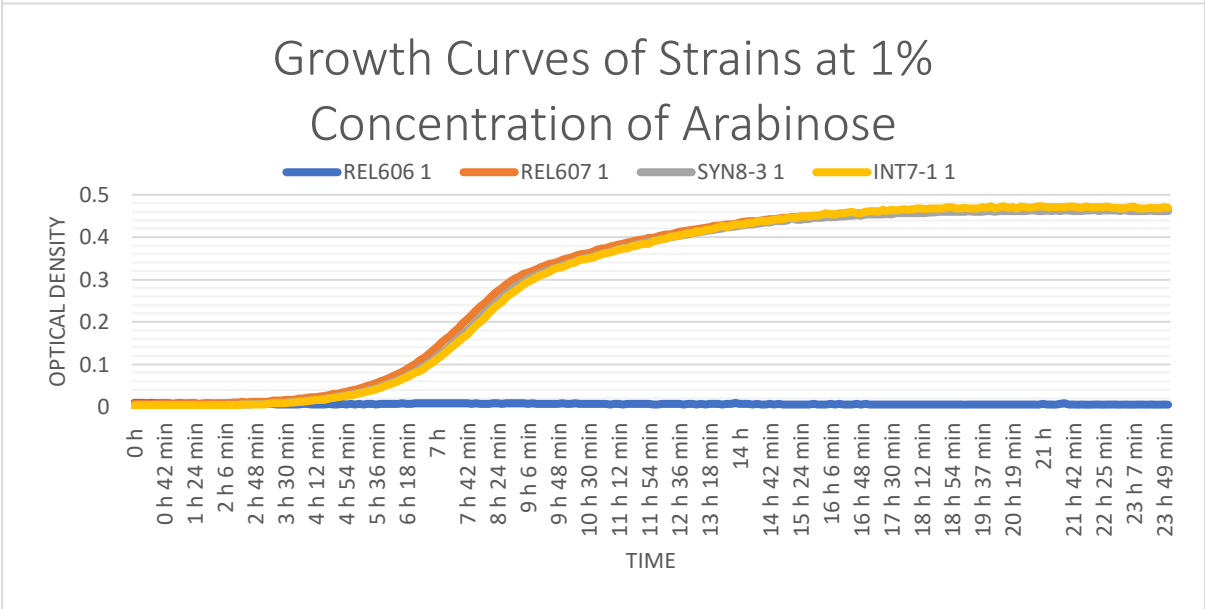
A)


Growth Curves of Strains at 0.02% Concentration of Arabinose
REL606 0.02
REL607 0.02
SYN8-3 0.02
INT7-1 0.02
OPTICAL DENSITY
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0 h
0 h 42 min
1 h 24 min
2 h 6 min
2 h 48 min
3 h 30 min
4 h 12 min
4 h 54 min
5 h 36 min
6 h 18 min
7 h
7 h 42 min
8 h 24 min
9 h 6 min
9 h 48 min
10 h 30 min
11 h 12 min
11 h 54 min
12 h 36 min
13 h 18 min
14 h
14 h 42 min
15 h 24 min
16 h 6 min
16 h 48 min
17 h 30 min
18 h 12 min
18 h 54 min
19 h 37 min
20 h 19 min
21 h
21 h 42 min
22 h 25 min
23 h 7 min
23 h 49 min
TIME

B)


Growth Curves of Strains at 0.2% Concentration of Arabinose
REL606 0.2
REL607 0.2
SYN8-3 0.2
INT7-1 0.2
OPTICAL DENSITY
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 h
0 h 42 min
1 h 24 min
2 h 6 min
2 h 48 min
3 h 30 min
4 h 12 min
4 h 54 min
5 h 36 min
6 h 18 min
7 h
7 h 42 min
8 h 24 min
9 h 6 min
9 h 48 min
10 h 30 min
11 h 12 min
11 h 54 min
12 h 36 min
13 h 18 min
14 h
14 h 42 min
15 h 24 min
16 h 6 min
16 h 48 min
17 h 30 min
18 h 12 min
18 h 54 min
19 h 37 min
20 h 19 min
21 h
21 h 42 min
22 h 25 min
23 h 7 min
23 h 49 min
TIME

C)


Growth Curves of Strains at 1% Concentration of Arabinose
REL606 1
REL607 1
SYN8-3 1
INT7-1 1
OPTICAL DENSITY
0.5
0.4
0.3
0.2
0.1
0
0 h
0 h 42 min
1 h 24 min
2 h 6 min
2 h 48 min
3 h 30 min
4 h 12 min
4 h 54 min
5 h 36 min
6 h 18 min
7 h
7 h 42 min
8 h 24 min
9 h 6 min
9 h 48 min
10 h 30 min
11 h 12 min
11 h 54 min
12 h 36 min
13 h 18 min
14 h
14 h 42 min
15 h 24 min
16 h 6 min
16 h 48 min
17 h 30 min
18 h 12 min
18 h 54 min
19 h 37 min
20 h 19 min
21 h
21 h 42 min
22 h 25 min
23 h 7 min
23 h 49 min
TIME

D)

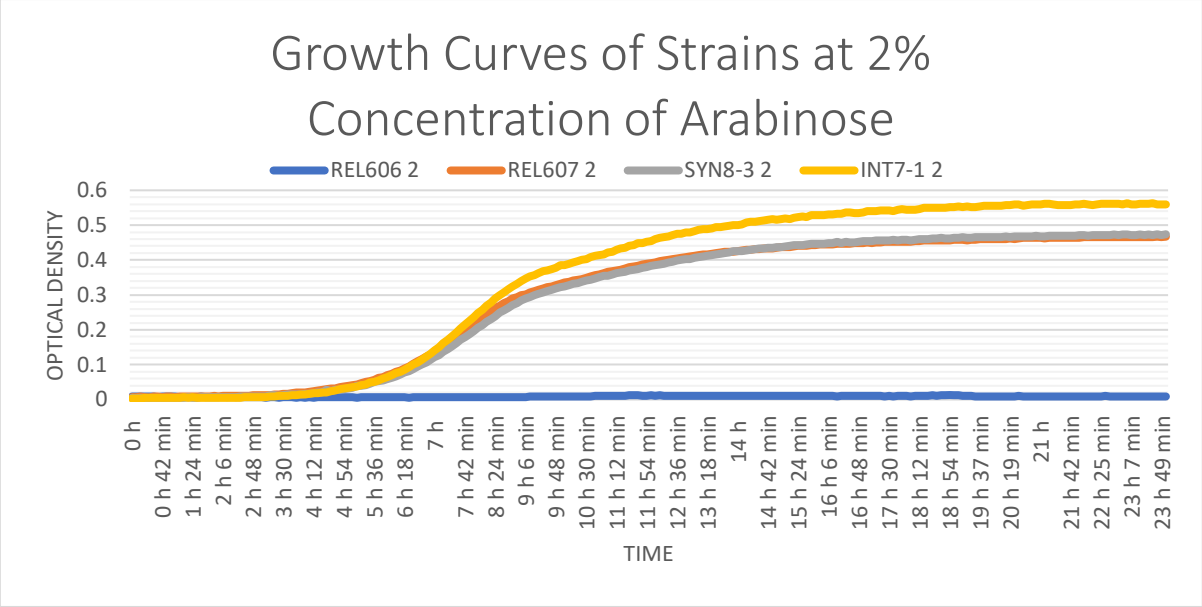
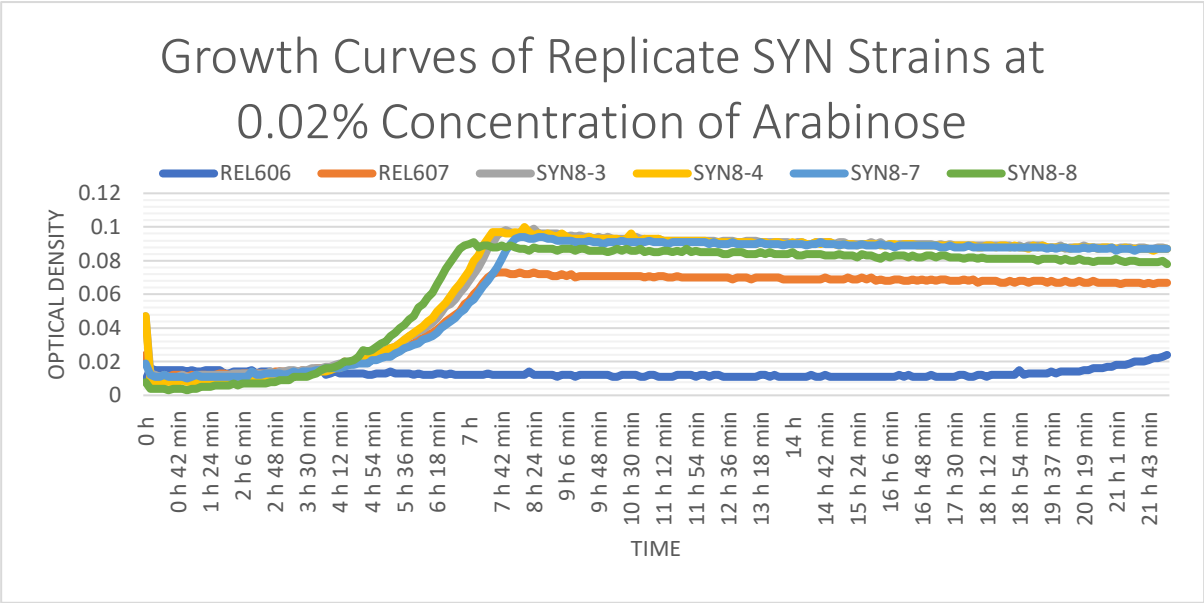
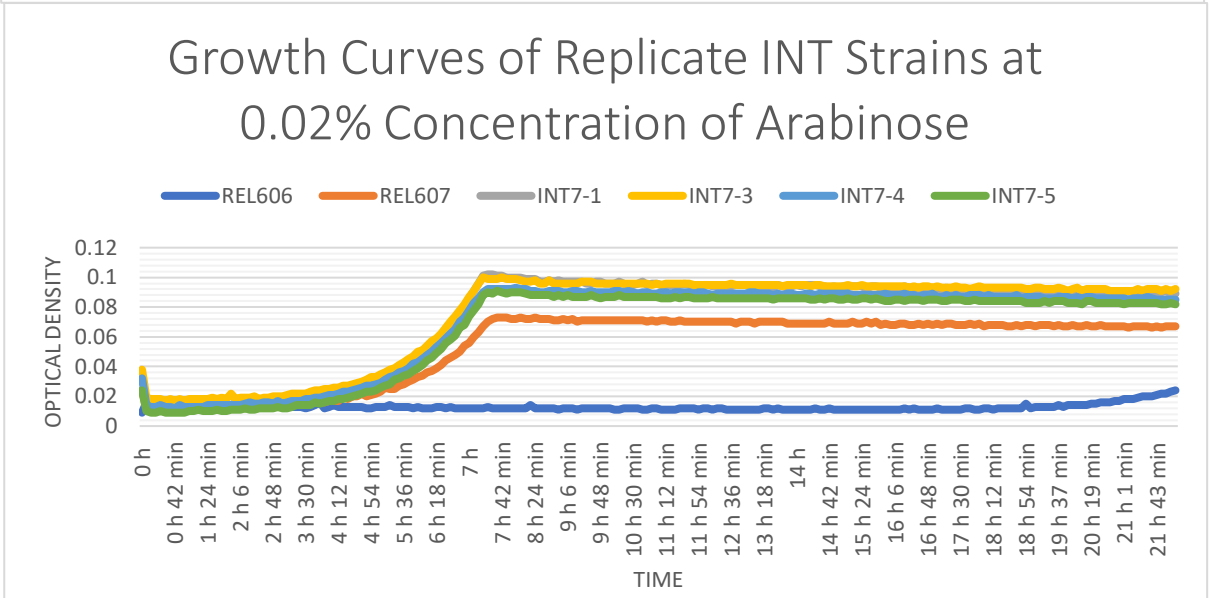


Figure 3.7. Four strains REL606, REL607, REL607::SYN and REL607::INT were grown in minimal media at four different concentrations of arabinose (0.02%, 0.2%, 1% and 2%). **A)** Growth curves of the two variant *araC* strains and two control strains, positive (REL607, grey) and negative (REL606, orange), cultured in minimal media supplemented with 0.02% arabinose. Growth rates appear not to differ much between strains. **B)** Growth curves of the two variant *araC* strains and two control strains, positive (REL607, orange) and negative (REL606, blue), cultured in minimal media supplemented with 0.2% arabinose. Growth rates appear show no variation between strains. **C)** Growth curves of the two variant *araC* strains and two control strains, positive (REL607, orange) and negative (REL606, blue), cultured in minimal media supplemented with 1% arabinose. Growth rates appear show no variation between strains. **D)** Growth curves of the two variant *araC* strains and two control strains, positive (REL607, grey) and negative (REL606, orange), cultured in minimal media supplemented with 0.02% arabinose. Growth rates show minimal variation between strains.

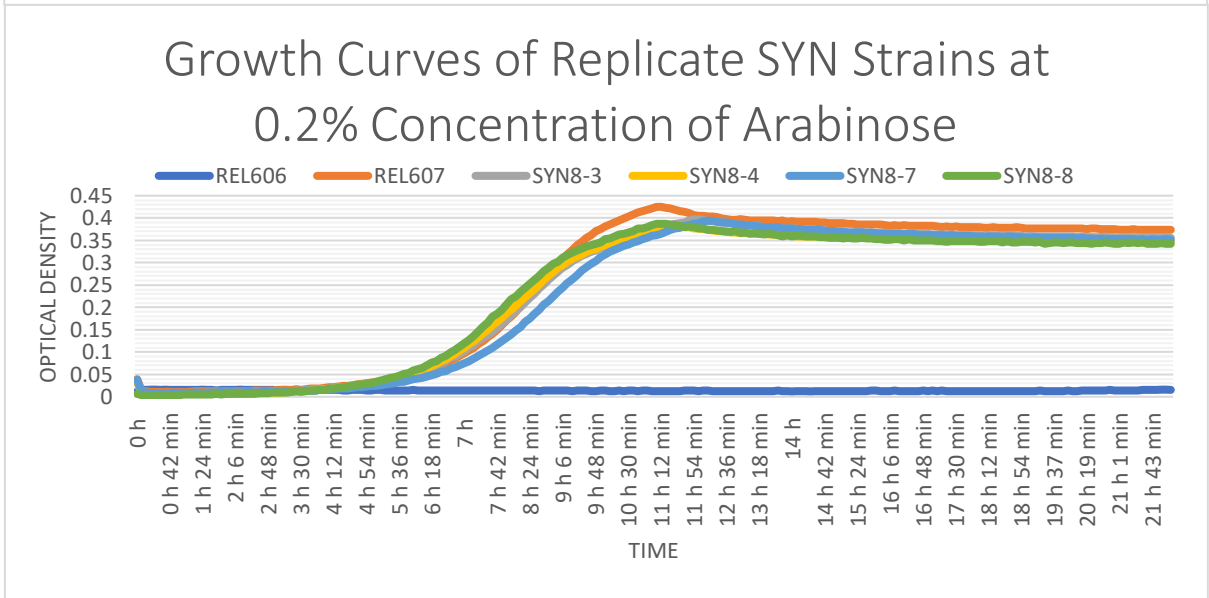
A)


Growth Curves of Replicate SYN Strains at 0.02% Concentration of Arabinose
REL606
REL607
SYN8-3
SYN8-4
SYN8-7
SYN8-8
OPTICAL DENSITY
0.12
0.1
0.08
0.06
0.04
0.02
0
0 h
0 h 42 min
1 h 24 min
2 h 6 min
2 h 48 min
3 h 30 min
4 h 12 min
4 h 54 min
5 h 36 min
6 h 18 min
7 h
7 h 42 min
8 h 24 min
9 h 6 min
9 h 48 min
10 h 30 min
11 h 12 min
11 h 54 min
12 h 36 min
13 h 18 min
14 h
14 h 42 min
15 h 24 min
16 h 6 min
16 h 48 min
17 h 30 min
18 h 12 min
18 h 54 min
19 h 37 min
20 h 19 min
21 h 1 min
21 h 43 min
TIME

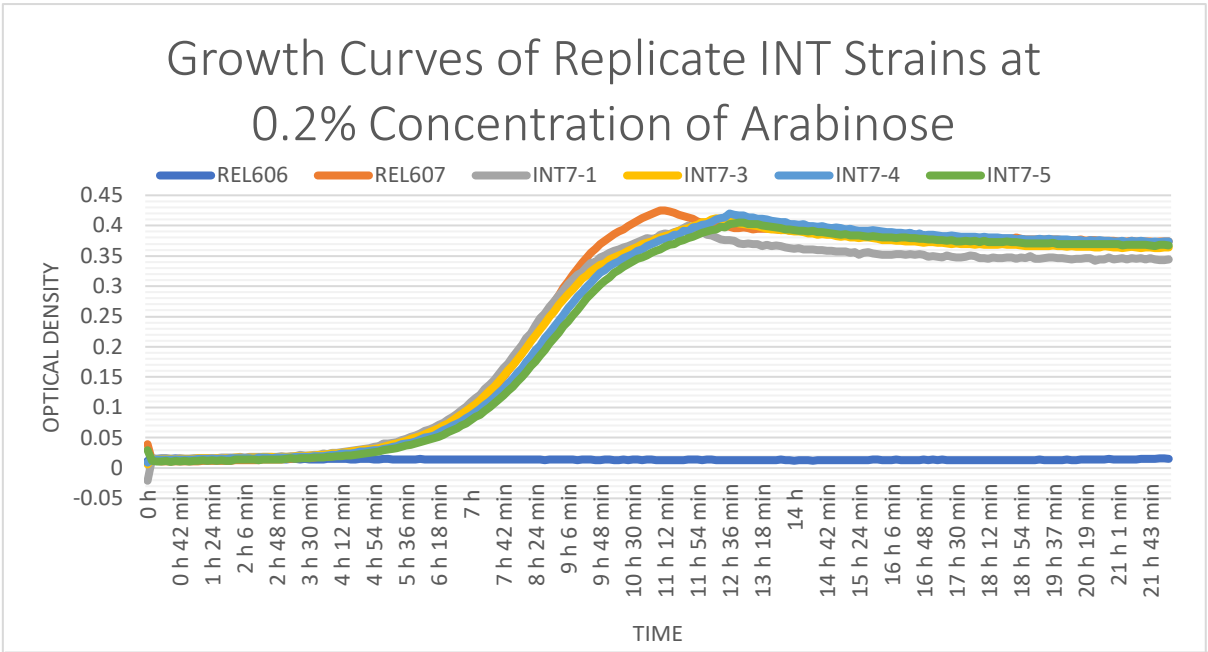
B)


Growth Curves of Replicate INT Strains at 0.02% Concentration of Arabinose
REL606
REL607
INT7-1
INT7-3
INT7-4
INT7-5
OPTICAL DENSITY
0.12
0.1
0.08
0.06
0.04
0.02
0
0 h
0 h 42 min
1 h 24 min
2 h 6 min
2 h 48 min
3 h 30 min
4 h 12 min
4 h 54 min
5 h 36 min
6 h 18 min
7 h
7 h 42 min
8 h 24 min
9 h 6 min
9 h 48 min
10 h 30 min
11 h 12 min
11 h 54 min
12 h 36 min
13 h 18 min
14 h
14 h 42 min
15 h 24 min
16 h 6 min
16 h 48 min
17 h 30 min
18 h 12 min
18 h 54 min
19 h 37 min
20 h 19 min
21 h 1 min
21 h 43 min
TIME

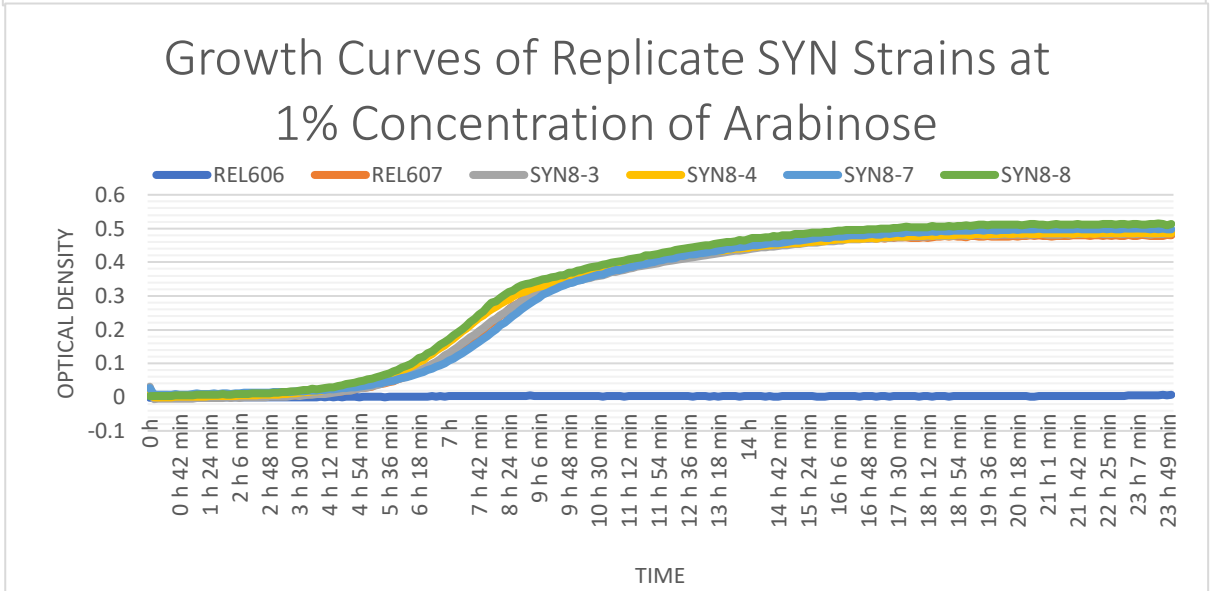
C)


Growth Curves of Replicate SYN Strains at 0.2% Concentration of Arabinose
REL606
REL607
SYN8-3
SYN8-4
SYN8-7
SYN8-8
OPTICAL DENSITY
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 h
0 h 42 min
1 h 24 min
2 h 6 min
2 h 48 min
3 h 30 min
4 h 12 min
4 h 54 min
5 h 36 min
6 h 18 min
7 h
7 h 42 min
8 h 24 min
9 h 6 min
9 h 48 min
10 h 30 min
11 h 12 min
11 h 54 min
12 h 36 min
13 h 18 min
14 h
14 h 42 min
15 h 24 min
16 h 6 min
16 h 48 min
17 h 30 min
18 h 12 min
18 h 54 min
19 h 37 min
20 h 19 min
21 h 1 min
21 h 43 min
TIME

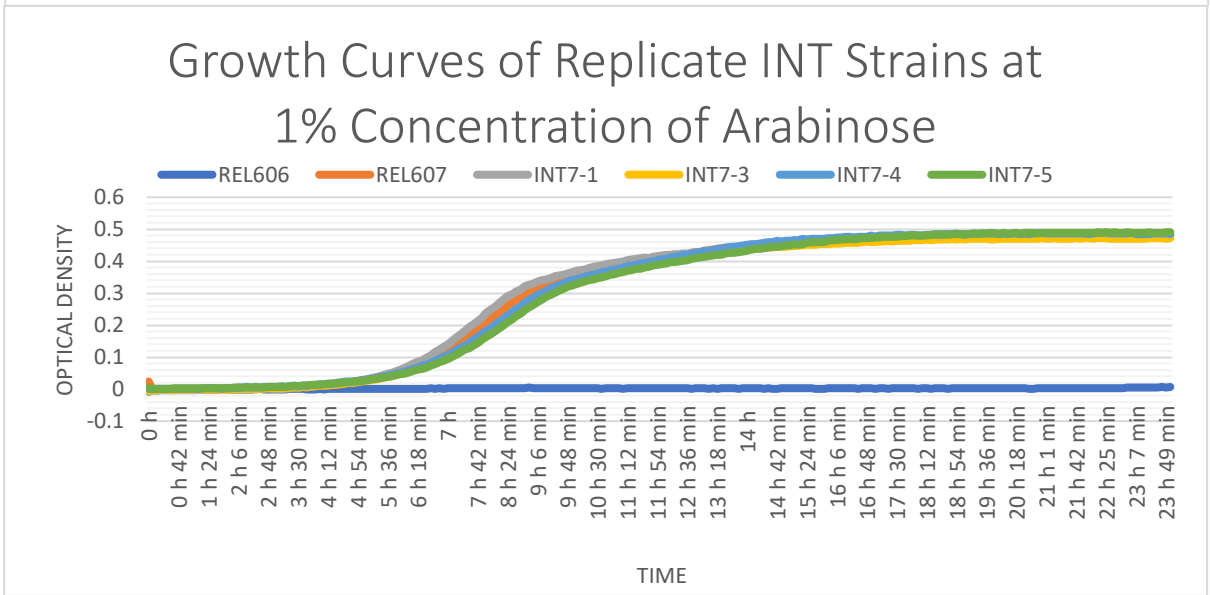
D)


Growth Curves of Replicate INT Strains at 0.2% Concentration of Arabinose
REL606
REL607
INT7-1
INT7-3
INT7-4
INT7-5
OPTICAL DENSITY
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
0 h
0 h 42 min
1 h 24 min
2 h 6 min
2 h 48 min
3 h 30 min
4 h 12 min
4 h 54 min
5 h 36 min
6 h 18 min
7 h
7 h 42 min
8 h 24 min
9 h 6 min
9 h 48 min
10 h 30 min
11 h 12 min
11 h 54 min
12 h 36 min
13 h 18 min
14 h
14 h 42 min
15 h 24 min
16 h 6 min
16 h 48 min
17 h 30 min
18 h 12 min
18 h 54 min
19 h 37 min
20 h 19 min
21 h 1 min
21 h 43 min
TIME

E)


Growth Curves of Replicate SYN Strains at 1% Concentration of Arabinose
REL606
REL607
SYN8-3
SYN8-4
SYN8-7
SYN8-8
OPTICAL DENSITY
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
0 h
0 h 42 min
1 h 24 min
2 h 6 min
2 h 48 min
3 h 30 min
4 h 12 min
4 h 54 min
5 h 36 min
6 h 18 min
7 h
7 h 42 min
8 h 24 min
9 h 6 min
9 h 48 min
10 h 30 min
11 h 12 min
11 h 54 min
12 h 36 min
13 h 18 min
14 h
14 h 42 min
15 h 24 min
16 h 6 min
16 h 48 min
17 h 30 min
18 h 12 min
18 h 54 min
19 h 36 min
20 h 18 min
21 h 1 min
21 h 42 min
22 h 25 min
23 h 7 min
23 h 49 min
TIME

F)


Growth Curves of Replicate INT Strains at 1% Concentration of Arabinose
REL606
REL607
INT7-1
INT7-3
INT7-4
INT7-5
OPTICAL DENSITY
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
0 h
0 h 42 min
1 h 24 min
2 h 6 min
2 h 48 min
3 h 30 min
4 h 12 min
4 h 54 min
5 h 36 min
6 h 18 min
7 h
7 h 42 min
8 h 24 min
9 h 6 min
9 h 48 min
10 h 30 min
11 h 12 min
11 h 54 min
12 h 36 min
13 h 18 min
14 h
14 h 42 min
15 h 24 min
16 h 6 min
16 h 48 min
17 h 30 min
18 h 12 min
18 h 54 min
19 h 36 min
20 h 18 min
21 h 1 min
21 h 42 min
22 h 25 min
23 h 7 min
23 h 49 min
TIME

G)

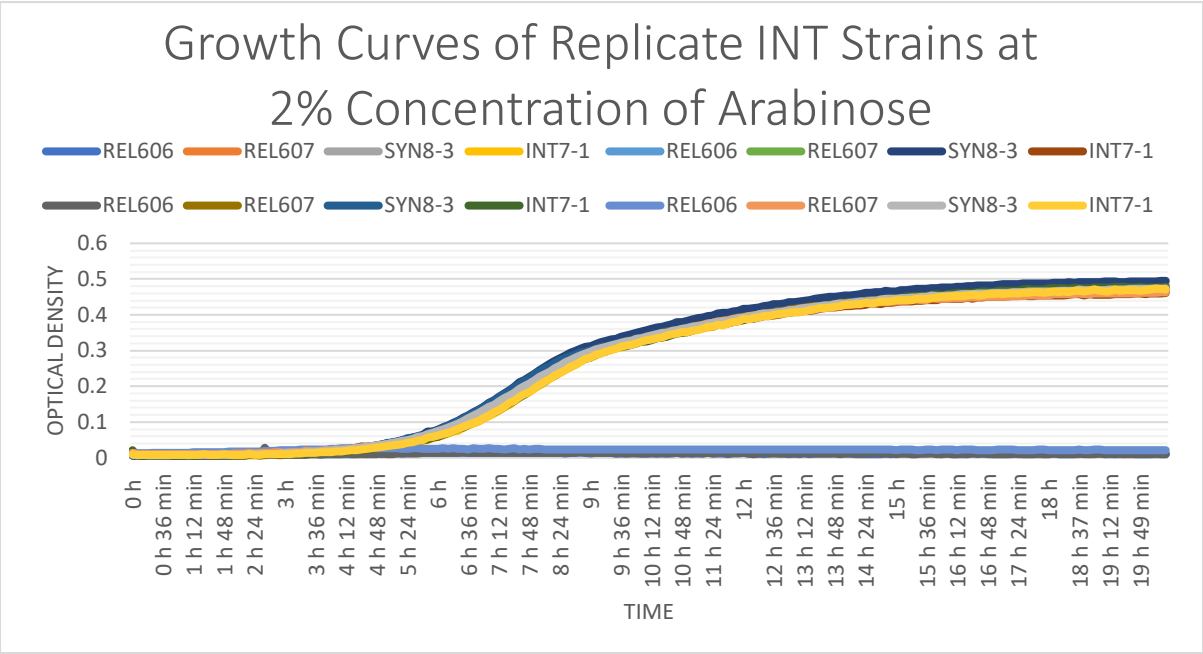


Figure 3.8. Replicate strains of REL607-SYN and REL607-INT were grown in minimal media at four different concentrations of arabinose (0.02%, 0.2%, 1% and 2%). **A)** Growth curves of replicate SYN strains and two control strains, positive (REL607, orange) and negative (REL606, blue), cultured in minimal media supplemented with 0.02% arabinose. Growth rates show some variation between strains. **B)** Growth curves of replicate INT strains and two control strains, positive (REL607, orange) and negative (REL606, blue), cultured in minimal media supplemented with 0.02% arabinose. Growth rates show some variation between strains. **C)** Growth curves of replicate SYN strains and two control strains, positive (REL607, orange) and negative (REL606, blue), cultured in minimal media supplemented with 0.2% arabinose. Growth rates appear to differ only in time to reach stationary phase. **D)** Growth curves of replicate INT strains and two control strains, positive (REL607, orange) and negative (REL606, blue), cultured in minimal media supplemented with 0.2% arabinose. Growth rates appear to differ only in time to reach stationary phase. **E)** Growth curves of replicate SYN strains and two control strains, positive (REL607, orange) and negative (REL606, blue), cultured in

minimal media supplemented with 1% arabinose. Growth rates show very little variation between strains. **F)** Growth curves of replicate INT strains and two control strains, positive (REL607, orange) and negative (REL606, blue), cultured in minimal media supplemented with 1% arabinose. Growth rates show very little variation between strains. **G)** Growth curves of four replicates of SYN 8-3 and 4 INT 7-1 strains and four control strains, positive (REL607, orange/brown) and negative (REL606, blue/dark grey), cultured in minimal media supplemented with 2% arabinose. Growth rates appear not to differ much between strains.

Lag Times Reveal No Differences in Initial Growth Rates Between *araC* Variant Strains and REL607

To determine whether there were initial deleterious fitness effects resulting from the synonymous changes to *araC* we compared the lag times between each of the *araC* variant strains and the wild-type REL607 control strain. Analysis of the growth curves indicated minimal difference in lag times between strains (Figure 3.7 and 3.8). Multiple students t-tests were performed to compare lag times. Comparing REL607-INT lag times with REL607 revealed no significant between the two strains at a 2% concentration of arabinose (t-test, $p = 0.238$). At this concentration, the comparison between REL607-SYN and REL607 was also not significant (t-test, $p = 0.204$). Additional t-tests were performed comparing strains at each concentration (0.02%, 0.2% and 1%), however no significant result was determined ($p > .05$). Comparisons between lag times of the designed strains, REL607-SYN and REL607-INT, were also not significant at each concentration level measured (t-test, $p > .05$)

Chapter 4
Discussion and Future Directions

Summary

The aim of this study was to further explore the mRNA: ncRNA avoidance model proposed by Umu et al. (2016) to elucidate how highly expressed mRNAs may have evolved to avoid unintended interactions with native ncRNAs and in-turn increase protein expression levels. This was explored by designing specific gene features that increase the potential for the mRNA to interact with ncRNAs. The arabinose metabolising gene of *E. coli* B strain REL606, *araC*, was modified for lower (*araC*SYN) and intermediate (*araC*INT) avoidance MFEs, which indicates binding affinity with ncRNAs, with respect to the wild-type. The modified *araC* genes were knocked-in to two separate lines of REL607, replacing the native *araC*. By continually growing these lines in minimal media supplemented with arabinose, where optimal expression of *araC* is highly important for cell fitness, it was hypothesised that natural selection would drive the evolution of avoidance between interactions of the designed *araC* mRNA variants and ncRNAs in *E. coli*.

The results presented in this thesis indicate that the alterations made to the 5' CDS of the *araC* gene were not sufficient to reduce their translation into protein via mRNA: ncRNA hybridization. Knocking in variant copies of the *araC* mRNAs with varying potentials for interaction with ncRNAs in *E. coli* B strain REL607, revealed no initial differences in growth between the designed strains and the wild-type control when cultured in Davis minimal media supplemented with arabinose. This result was not hypothesised at the outset of this research.

We had originally hypothesised that the synonymous changes made within the first 21 nucleotides of the *araC* gene would increase the total number of interactions the *araC* mRNA would have with native ncRNAs in *E. coli*, thus inhibiting mRNAs from being translated by the ribosome and subsequently knocking down the expression of the *araC* protein. The growth assays involved growing strains REL607-SYN and REL607-INT and REL607 at several different concentrations of arabinose in a 24-well plate to determine which concentration would produce the greatest differences in growth between strains. However, the only observed growth differences were observed between the different arabinose concentrations with no significant difference in growth rate between the strains at each concentration (see Chapter 3, pg. 58). Specifically, there was no discernible difference between the *araC* variant strains, REL607-SYN and REL607-INT, and the wild type REL607. Growth experiments were replicated in minimal media comparing the strains in a 24-well plate at only one or two different arabinose concentrations. The variant strains grown at two concentrations in a single plate (0.02%, 0.2% or 0.2%, 1%) showed no significant difference in growth rates between the strains at each concentration level (see Chapter 3, pg. 58). These results indicated that there is no detectable effect of the synonymous changes in *araC* on the growth of the variant strains in minimal + arabinose media. Further experiments attempted to determine whether the strains being used to assay growth were expressing the correct constructs. We measured growth between replicates of each of the designed strains, REL607-SYN, REL607-INT and the wild-type control REL607, however no differences in growth were detected. Further to this Sanger sequencing following these growth experiments confirmed that the a*raC* variant REL607 strains were indeed expressing the variant copies of *araC*. In addition to these growth differences determined using growth rates or doubling times, differences in growth were also determined using lag times. The computed lag times of REL607-SYN, REL607-INT and the

control, REL607, were compared at each concentration of arabinose (0.02%, 0.2% or 0.2%, 1%). At each concentration, the lag times between the variant *araC* strains and wild type were shown to be not significantly different from one another (see Chapter 3, pg. 65).

Modifying genes for varying interaction potentials with ncRNAs has previously been shown to be a good indicator for the level protein expression (Umu et al., 2016). Interestingly, we found that altering the 5' end of the *araC* gene, a region which has been shown to be highly significant for avoidance, produced no discernible fitness effects between the designed strains and the wild type control strain. While the basic premise behind these two experiments is the same there are several differences between the methodologies used to test the mRNA: ncRNA avoidance model, the main difference being the gene that was used to assay RNA-RNA interaction. The GFP gene was taken from a foreign host and transplanted into *E. coli*, whereas the *araC* gene is already native to *E. coli*. An important distinction here is that *araC* has evolved in parallel with native ncRNAs in *E. coli*, while the GFP is being expressed in a non-native context. This suggests to me that simulating avoidance of RNA-RNA interactions in a native context is more difficult than in a non-native context. It has previously been established that there is a reduced capacity for native mRNA: ncRNA interactions (Umu et al., 2016). The designed *araC* variants each differed from the wild type by a total of 4 nucleotides. Therefore, how avoidance is facilitated may be more complex than simply changing a few nucleotides that limit the degree of interaction between RNA pairs. This research therefore provides an interesting test of whether altering the nucleic acid sequence of a native gene can reduce its established avoidance potential. In the following sections I will discuss the results and limitations of this research as well as directions and the implications for further research.

Addressing Statistical Noise in Assaying Growth Differences

The variation in the level of significance between replicates of the INT strain when determining differences in growth using growth rates or doubling time (see Chapter 3, pg. 58-59) speaks to the limitation of the statistical package that was used, GrowthRates. The following is an excerpt from the GRplot documentation, which is a program that is used to help trouble-shoot the results of running GrowthRates:

"The program GrowthRates is by no means perfect. Sometimes the rates it reports don't seem to make sense; sometime the time points used to estimate the growth rates are within the lag period; sometimes the correlation coefficient (R) is low (<0.995). When that occurs the GrowthRates documentation suggests plotting the ln(O.D.) vs time for the offending well to help understand what is going on." – GRplot Documentation, January 13, 2018.

The issues presented here seem to align with some of the results that I obtained using the GrowthRates package. The reported doubling times for the negative control strain REL606 were often very short, 2-3 minutes, despite this strain reaching only very low maximum optical densities. GrowthRates determines the growth rates of strains by considering a window of five time-points and calculates the slope of optical density (OD) vs time and saves that value. It then moves one time point and considers the next window of five and saves the product of the slope and the correlation coefficient, R. After it has calculated all 5-point slopes and their coefficients up through the highest OD it uses the time points set whose slope x R product was highest to determine the initial time points from which the growth rate will be determined. Using this method strains that grow poorly may subsequently have fast growth rates calculate, as OD in the beginning stages of culturing can quickly double but not increase

much beyond very low densities. As an example, a strain may only reach a low OD and still demonstrate fast growth rates. Measurements of REL606 in the same plate, in different wells under the same arabinose concentration (2%) calculated doubling times of 19.2 minutes, where maximum OD was 0.028 and 542.7 minutes where maximum OD was 0.017. Thus, growth rate using this package is largely dependent on the window of 5 time-points.

The Level of Selection on *araC* Gene

In Chapter 1 I discussed the effects of codon usage bias and mRNA secondary structure on the expression of protein. The main take away from these sections is that both mRNA secondary structure and codon usage have a significant impact on protein expression, especially in the 5' region. However, the degree to which these phenomena impact protein expression does not completely account for the discrepancy between mRNA and protein abundances.

Explanations of codon usage patterns within and between species fall into two distinct categories: mutational based and natural selection based. Mutational explanations suggest that codon biases arise from mutational processes such as point mutations and biases in repair mechanisms. These explanations are neutral as they offer no fitness advantage or detriment for synonymous mutations. Conversely selection based hypotheses suggest that synonymous mutations influence the fitness of an organism and as a result can be fixed or lost through evolution (Plotkin & Kudla, 2011). If mRNA: ncRNA avoidance evolved through selection for genetic mutations that decrease the ability for mRNAs to hybridize with ncRNAs then we would expect that a gene carrying codons designed for high interaction with ncRNAs

in an environment where optimal expression of the gene is necessary for growth would quickly evolve by maintaining mutations that limit these interactions. A common explanation for variation in codon bias across a genome is selection (Plotkin & Kudla, 2011). Codon bias in the *E. coli* genome has been established to be more extreme in highly expressed genes to match the skew in iso-accepting tRNAs, thus providing a fitness advantage via increased translation efficiency of protein synthesis (Ikemura, 1981). The *araC* gene of *E. coli* is expressed at low background levels when not induced, and only undergoes high levels of expression when it encounters significant levels of inducer (Siegele & Hu, 1997). As such *araC* does not demonstrate strong codon adaptation (Sharp & Li, 1987) compared with other genes in *E. coli* which subsequently suggests that codon usage is not under strong selection. If *araC* is not under strong selection for codon bias then I would also expect it to minimally select for avoidance of potentially deleterious interactions between the mRNA and native ncRNAs. The growth rates we see in *E. coli* may therefore be indicative of a very basal level of expression when encountering unintentional interactions with ncRNAs. Perhaps by turning the dial the other way and optimising *araC* for high avoidance we may see significant differences in growth relative to the wild type. If this is true with then the experimental set up we have implemented *araC* is a poor candidate gene for assessing the evolution of avoidance. The *araC* gene was selected as our gene of interest given its vital role as the regulator of the arabinose operon (Schleif, 2010). However, upon measuring the native avoidance of the additional structural genes of the arabinose operon, *araBAD*, it was found that these genes have a much higher avoidance MFE in their first 21 nucleotides (see Chapter 2, pg. 39), suggesting that these genes have a more developed avoidance profile. Presumably genes that have a higher avoidance MFE have evolved through selection to minimally interact with ncRNAs to a greater degree than genes with lower avoidance MFE values. This suggests that

these genes may exhibit greater ramifications of sequence alterations that would increase their potential for interactions between the gene transcript and native ncRNAs than *araC* and as such would provide a better and more easily observable initial detriment in fitness for assaying growth and the impact of stochastic RNA-RNA interactions. It would be interesting to determine whether the native avoidance of araB, araA and araD is correlated with expression.

Caveats with Predicting RNA-RNA Interactions using MFE

It is worth discussing the process used to the produce the synonymously variant *araC* genes to be knocked into *E. coli* to demonstrate the evolution of avoidance. The designed variants were generated using the same method utilised by Umu et al. (2016). This process utilised an algorithm to determine the interaction potential between generated *araC* gene variants and a set of established ncRNAs in *E. coli* B strain REL606 based on minimum free energy (MFE) values. MFE methods are popular in the prediction of RNA-RNA interactions (Lai & Meyer, 2016; Lorenz et al., 2011; Pain et al., 2015). However, this method is not without limitations. In particular RNA-RNA interaction prediction algorithms that utilise MFE methods are not always accurate (Dieterich & Stadler, 2013; Lai & Meyer, 2016; Pain et al., 2015). In a comprehensive benchmark of 15 freely available RNA-RNA interaction prediction tools where correct interactions were known, it was found that 15 of 154 RNA interaction pairs could not be correctly predicted by any of the algorithms (Umu & Gardner, 2016). Prediction of mRNA-sRNA interactions in *E. coli* using trusted, experimentally validated sRNA/target pairs was assessed in another study using several RNA target prediction methods (Pain et al., 2015). Five prediction programs, IntaRNA/CopraRNA, RNAplex, TargetRNA2 and RNAup, were evaluated based on speed and ability to identify the true interaction regions between

sRNA/target pairs. The degree of accuracy between these tools varied from 56-79% (Pain et al., 2015). While MFE methods are widely used, these studies demonstrate that a large proportion of true interaction pairs remains unaccounted for in some data sets. As such there is the potential that the predicted avoidance MFE values of our designed variants are inaccurate. This would suggest that some proportion of our designed variants may have estimated interaction potentials that are based on inaccurate pairings. Thus, our designed *araC* variants and native *E. coli* ncRNAs may be less likely to interact than expected. Upon generating the initial *araC* variants the first 21 nucleotides of two specific variants with low and intermediate avoidance values relative to the wild-type were extracted and sent away for synthesis (Macrogen).

Contrasting Synonymous GFP and *araC* mRNAs

The final *araC* constructs had synonymous changes at four sites within the first 21-nucleotide region resulting in a change of 4 of the 7 codons in the genes start site relative to wild type *araC*. While the process of designing of the *araC* constructs is very like the design of the GFP constructs used by Umu et al. (2016), there are several key differences between the final variants used in each assay. The GFP constructs generated by Umu et al. (2016) were designed to capture the extremes of one variable while controlling for the others. The design for their avoidance constructs was optimised for high or low avoidance in the first 21 nucleotides of the GFP CDS, near average mRNA secondary structure in the first 37 nucleotides and near average codon adaptation. As such the constructs that were used in the GFP experiment were entirely computer generated. In the case of this study, only the initial 21 nucleotides of the *araC* CDS were computer generated, while the remainder of the sequence was identical to *araC* gene taken from the genome of REL606 and thus more closely represented the wild-

type sequence (Figure 4.1). In addition, the design script allows for variation at sites other than the 3’ codon position as well as positions upstream of the first 21 nucleotides of the gene. Given that other regions along the mRNAs of the synonymously variant GFPs were identified as being of significance for avoidance (Umu et al., 2016), these additional sequence changes may facilitate binding at regions other than the 5' end of the mRNA thus allow for RNA-RNA interaction at multiple sites along the transcript. In future experiments, it would be interesting to test the ability for an entirely computer generated *araC* variant to interact with ncRNAs in *E. coli*. This would indicate the importance of sequence variation at other regions along the gene that make the mRNA more susceptible to hybridisation with ncRNAs.

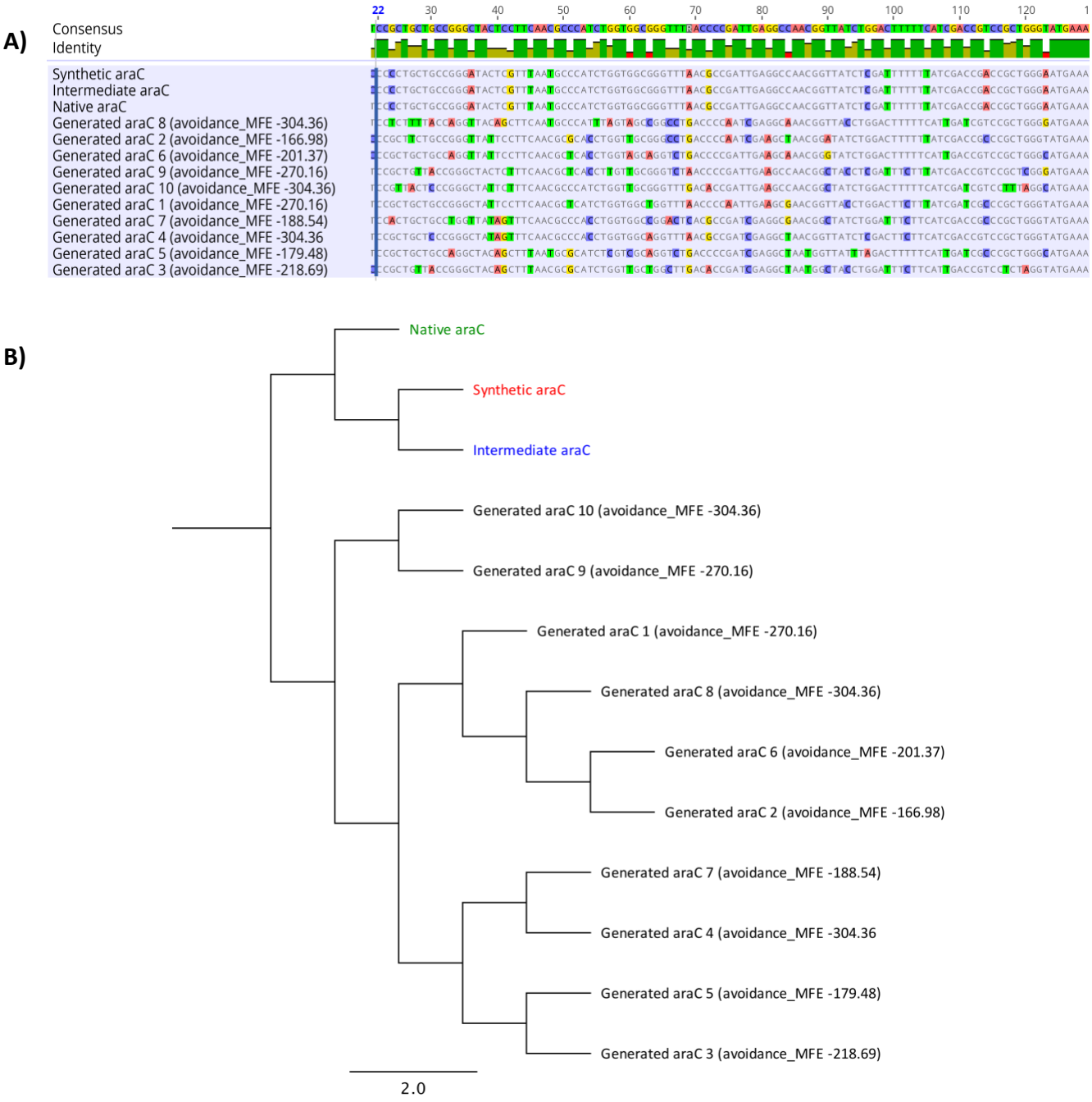


Figure 4.1. **A)** Multiple alignment of synonymously variant *araC* genes that were generated using the avoidance script and the native *araC*. This alignment reveals the differences in nucleotides composition of the araC variants used in the final avoidance assay and those generated using the avoidance script. The alignment was created in Geneious using MUSCLE **B)** Gene tree of synonymous *araC* variants. 10 of the synonymously variant *araC* genes that were generated using the avoidance script were randomly selected along with the two *araC*

variants used for the allelic replacement protocol and the native *araC*. The tree was generated in Geneious (version 10.2.3) the alignment type used was a global alignment with free end gaps and a cost matrix of 65% similarity. Tamura-Nei was used to compute genetic distances, neighbour-joining was the tree building method.

Is *araC* Robust to Synonymous Mutations?

Another indication that *araC* may have been a poor choice for this experiment is that expression of *araC* may be robust to synonymous sequence alterations. This was first hinted at in a study which revealed that between two strains of *E. coli* the *araC* gene differed by more than 9 nucleotides. These changes all occurred in the third wobble position, none of which produced a change in the amino acid sequence of the protein (Stoner & Schleif, 1982). This study suggested that conservation of the amino acid sequence of *araC* was of selected for while there was no selection to maintain the nucleotide sequence. After we began this project a study considering the variation in mutational robustness between different classes of proteins found that introducing random synonymous mutations in *araC* did not reduce fitness in *Salmonella typhimurium* (Lind, Arvidsson, Berg, & Andersson, 2017). 18 synonymous mutations were introduced at positions across the *araC* gene. The mutations were transduced into strains with fluorescent protein marker cassettes inserted at neutral positions and placed in direct competition with the wild-type control strain. Neutrality of these mutations was determined based on a selection coefficient between -0.004 and 0.004. It was also found that additional genes in the arabinose operon, *araD* and *araE* also showed no fitness effects due to synonymous mutation. These three genes specify three different classes of protein, transcription factor (*AraC*), enzyme (*AraD*) and transporter (*AraE*). In total forty-seven synonymous substitutions were introduced across all three arabinose metabolism genes and

all of them were classified as “neutral” which here means having no discernible difference from the wild type. These findings are consistent with the results of this study which found no distinction between the growth rate of *E. coli* strains that had synonymous changes introduced in the first 21 nucleotides of *araC* and the wild type control. While these experiments were conducted using different bacterial species the two exhibit highly similar genome content and metabolic networks (Sargo et al., 2015). This is not to say that synonymous changes do not impact organismal fitness. The same study also assessed the robustness of ribosomal proteins to synonymous mutation. They found that the large majority of synonymous mutations were deleterious, with only 2 out of 38 mutations being classified as neutral for ribosomal protein genes. This suggests classes of protein that are under constant selection are less robust to synonymous mutation than proteins that only under selection for short periods. One important distinction to be made between these two studies however is that none of these synonymous mutations produced by Lind et al, (2017) occurred in the first 21 nucleotides, the region demonstrated as being highly important for avoidance as it contains the translation initiation site, the rate limiting step for translation. The first synonymous mutation is introduced at least 50 nucleotides downstream of the start codon. These mutations also occur only once in each strain which would may subsequently reduce the number of interaction sites for ncRNAs. This limits these mutations from being implicated in having an impact on avoidance. This study however demonstrates the impact synonymous mutations can have on protein expression while also indicating that several genes of the arabinose operon are able to withstand multiple synonymous mutations without affecting expression.

Methods for Measuring Protein Expression

In chapter 3 I described the methodology that was used to produce the *araC* variant lines of REL607, REL607-SYN and REL607-INT. While this process was effective at producing a knock-in and replacement of the wild-type *araC* with our low and intermediate avoidance *araC* variants it was very time-consuming and we saw no impact of this knock-in on the growth of the bacteria under the "selective" conditions. While this outcome suggests that alterations to the 5' end of *araC* have no detectable impact on ncRNA interactions with the *araC* mRNA this is not definitive. It is possible that changes to this gene to make it more receptive to hybridisation with native ncRNAs in *E. coli* have an undetectable effect when assaying growth in minimal media supplemented with arabinose. Given this unexpected result it is possible that the phenotypic effect of altering *araC* is not detectable by simply measuring OD. This speaks to the limitations of our experimental approach. Firstly, we are not directly measuring protein levels by assaying OD, rather we are measuring cell density in culture and using it as a proxy for *AraC* protein levels. This could be improved for future experiments. Perhaps a better approach to quantifying protein levels would have been to utilise protein purification techniques. Affinity tags consisting of six polyhistidine residues are commonly used as a means of purification and subsequent quantification of the tagged protein. Histidine readily forms bonds with transition metal ions that have been immobilised to a column (Kimple, Brill, & Pasker, 2013) allowing proteins carrying these residues to be easily purified when biological samples are filtered through it. This process typically involves adding the tag to either the C or N terminus of the protein but optimal placement of the tag is protein specific so it is important to have a good understanding of the structural nature of the protein before incorporating a tag (Bornhorst & Falke, 2000).

How Sequencing Could Reveal Alternative Avoidance Mechanisms

Sanger sequencing of the *araC* gene in the REL607-SYN and REL607-INT using the araC -721bp/+142bp lines revealed that indeed the designed gene constructs were integrated at this site in the chromosome. However further sequencing may have elucidated more information. In Chapter 1 we predicted several different adaptive responses to the selective pressure imposed on these *araC* gene variant lines. For instance, we predicted that a mutation in the promoter region may have caused an increase in the overall transcription of the *araC* gene variants subsequently leading to a greater number of transcripts which would reduce the impact of avoidance. An interesting idea posited by Plotkin and Kudla (2011) is that if high protein levels are advantageous under strong selective pressure, from an evolutionary perspective, it would seem easier to tune a promoter for increased transcription than to select on hundreds of different SNPs, each of which would only marginally impact the overall expression of the promoter. Sequencing of the promoter region of *araC* therefore may have shed some light on the activity of the promoter in relation to the growth dynamics of these strains under the selective conditions. Additionally, we hypothesised that a duplication of the *araC* gene may have resulted in an adaptive response that would increase the amount of mRNA available for translation, thus increasing the overall expression. For this reason, whole genome sequencing of the designed strains following their introduction to the selective environment may have been useful in revealing such an event. However, while these adaptive responses are a possibility, it was expected that they would occur over the course of an evolution experiment. As such for these responses to selection to be considered as an explanation for the similarities in growth dynamics of the designed strains compared with the wild-type they would have had to have occurred over a very rapid time frame (24-hours), which would indicate that the selective pressure to express *araC* is extremely high. This kind

of rapid adaptation to selection is possible, but is perhaps unlikely. Further to this analysis, of the growth curves of the designed strains in comparison to the wild-type found that the lag times for the strains were not significantly dissimilar at each concentration of arabinose. In other words, both the designed strains and the wild-type appear to reach the exponential growth phase at the same time, suggesting that there is no initial detriment to having an altered *araC* start site. Additionally, for an adaptive response where growth rate is increased due to an increase in the number of available transcripts, transcription levels for these mRNAs would have to be extremely high to outcompete the abundance of ncRNAs that are able to stochastically bind them.

The Use of GFP mRNAs in a Non-Native Context

Umu et al., (2016) determined that native interactions between mRNAs and ncRNAs consistently have higher (less stable) free energies when compared to negative controls which indicated a reduced capacity for interaction among native RNAs. This finding led to the conception of the mRNA: ncRNA avoidance model. Following on from this research we hypothesised that RNA-RNA interactions are avoided through accumulation of mutations in the gene sequence, primarily in the first 21 nucleotides, that reduce the stochastic intermolecular binding potential of mRNAs with ncRNAs. As our approach to gene design replicated Umu et al. (2016), it is therefore necessary to illustrate the limitations in their experimental approach to determining the mechanism by which avoidance is facilitated. The impact of avoidance, or lack thereof, on protein expression was assessed by measuring the effects of synonymous changes to the first 21 nucleotides of GFP mRNAs in *E. coli* that would increase or decrease its interaction potential with native ncRNAs. In this case, the gene was moved from its original organism, the jellyfish *Aequorea Victoria* (Shimomura, 2005), into *E.*

coli BL21(DE3), synthetically designed to demonstrate high or low avoidance, and expressed from a plasmid, pET-32a, at high level. Protein expression was then measured with arguably low sensitivity using fluorescence, in an environment and at a temperature very different from where the gene had evolved, with no connection between the genes function and cell fitness. While this and other studies certainly contribute to our understanding of the mechanics of molecular biology, their ability to be applied to evolution or organismal fitness should be regarded as important but necessarily preliminary.

Future Directions

Designing an *AraC*-GFP Fusion Protein for Assessing mRNA:ncRNA Avoidance

Functional *araC* transcripts are essential for *E. coli* when arabinose is the sole carbon source provided in the culturing media. Our *araC* variants were designed with differing affinities for interaction with native ncRNAs in *E. coli*, however current efforts to assay fitness effects have not revealed any difference between the strains harbouring genes designed for different levels of avoidance. To assess whether there are any differences in protein production of the *AraC* protein from the *araC* genes with different predicted levels of avoidance future experiments could utilise an *AraC*-GFP fusion protein. Tagging proteins of interest with fluorescent proteins can be used to detect the level of a protein of the target gene, with the level of expression quantified by the amount of fluorescence emitted by the organism. Here I will outline several methods one could use to create a fusion protein of *AraC* and GFP.

Determine Regions that are Permissive for Fusing Proteins

Using random transposon insertion, it is possible to reveal permissive sites for protein insertion that may not be predicted based on structural or functional protein models (Sheridan et al., 2002). Often when producing fusion proteins, the approach is to tag either the N- or C-terminal of the protein with GFP. In this case, however, it will be important to carefully look at the protein structure and determine whether or not fusion with GFP will interfere with the activity of the *AraC* protein itself. As the strongest signal for avoidance is within the first 21bp of a gene, we have thus designed our *araC* transcripts with changes to the first 21bp, meaning that tagging the N-terminus could present issues during expression, as it may interfere with the predicted mRNA: ncRNA interaction region. *AraC* functions as a homodimer, meaning that the fusion protein should not interfere with dimerization. In the absence of arabinose, the *AraC* regulator protein represses expression of the *araBAD* operon by binding to two regions of the DNA and forming a loop structure. The loop represses expression of the operon itself by hindering the binding of RNA polymerase to the *pBAD* promoter (Dirla, Chien, & Schleif, 2009). It will be important to design this fusion protein in a way that does not interfere with protein function.

Construction of AraC:GFP Fusion Proteins with PCR (Overlap Extension PCR)

Provided that tagging the C-terminal of the protein is sufficient, construction of an *AraC*-GFP fusion protein should be relatively straightforward. This method involves performing two independent PCR reactions, with partially matching overhangs on the primers, followed by a third PCR reaction that will fuse the products of the first two PCRs to form one fusion product (Bryskin and Matsumura, 2010). When designing fusion proteins in this manner, for a C-terminal fusion, it is necessary to eliminate the start codon of GFP and the stop codon of *araC*

to prevent either sole expression of GFP, and/or a truncated fusion protein where GFP is absent. This method can then be used to either introduce the fusion protein onto a plasmid, which can subsequently be used to assay protein production directly or to introduce the fusion protein into the chromosome using methods already well-established in the Poole Lab. Figure 4.3 explains the general idea of this protocol. A PCR is first performed with the primers labelled F1 and R2. R2 is designed in a way that it excludes the stop codon of araC, as well of the start codon of GFP, but also spans the beginning of the GFP sequence. A separate PCR is also performed using the F3 and R4 primers. Like R2, F3 is also designed to exclude the start and stop codons while spanning both the araC and GFP sequences. A third PCR is then performed using primers that span the entire fusion sequence. The idea is that these two PCR products will ligate, forming the desire fusion product.

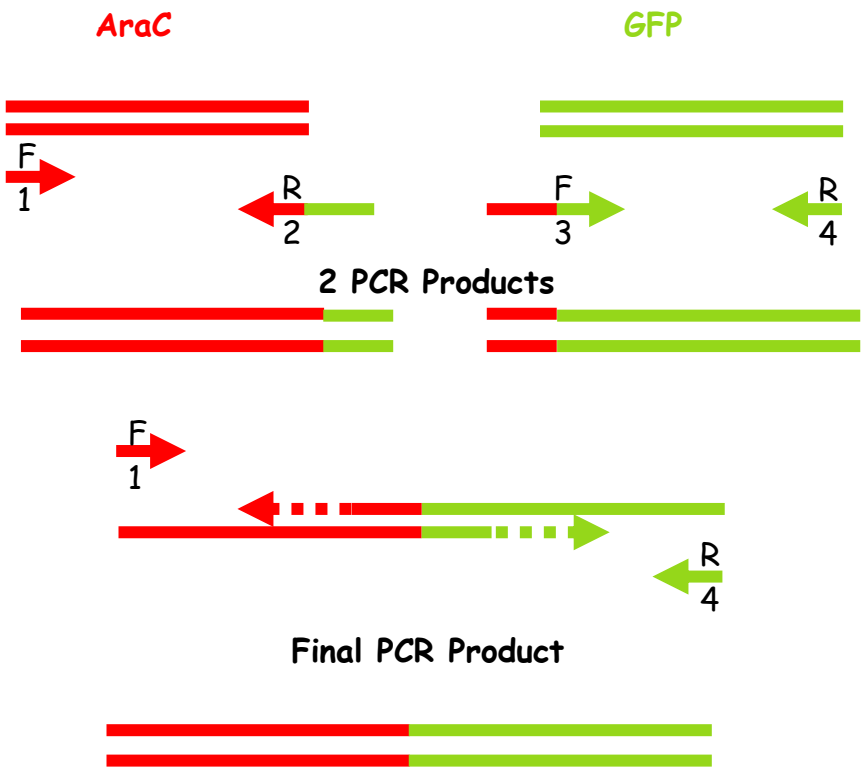


Figure 4.3. Outline for construction of an *araC*-GFP fusion protein. This process utilises overlap extension PCR that amplifies over a region of each gene (Bryksin & Matsumura, 2010). The amplified regions are then incorporated into a second PCR reaction that fuses the previous two PCR products together, resulting a gene product that carries a fluorescent marker. This allows one to easily determine the level of gene expression via fluorescence.

Gene Choice and Optimal Experimental Design

The choice of gene in the context of avoidance needs to be thoroughly researched to provide an accurate test of the hypothesis that mRNA: ncRNA avoidance evolved via mutations to the gene, and subsequently the mRNA, that minimise interaction with ncRNAs. Post experimental testing, research of *araC* has indicated that this gene, and others among the arabinose operon, may be robust to synonymous mutations, in that such sequence changes do not affect host cell fitness (Lind et al., 2017). The results of this thesis are in alignment with this. Given that synonymous changes to a gene that result in lower fitness of the host organisms could be the result of increased interactions with ncRNAs in future experiments it will be important to investigate the host fitness of any candidate genes in response to silent mutations, specifically candidate genes that are known to be negatively impacted by synonymous sequence changes. Further verification of the gene of choice could include a transcriptome analysis. Optimal candidate genes should reveal that synonymous mutations in the nucleotide sequence, resulting in lower fitness of the host organism, should not correlate with reduced levels of the candidate genes transcript. This would indicate that expression of the gene is impacted at the translation step, which increases the likelihood that expression is being hindered when the candidate mRNA hybridizes with native ncRNAs in the host. Alternatively, this could be estimated using reverse transcriptase quantitative PCR (RT-

qPCR). This involves the extraction of mRNA from the cell and subsequently reverse transcribing the mRNA into complementary DNA (cDNA). The cDNA can then be used as a template for qPCR using primers designed to amplify the target gene. As PCR is running fluorescent tags attached to the primers will give an estimate of the total concentration of candidate mRNAs. From this the starting concentration of the target mRNA can be determined. To summarise if mRNA: ncRNA avoidance evolves through mutations to the gene that limit transcribed mRNA from interacting with ncRNAs then a transcriptome analysis should reveal no impact on the number of transcripts produced.

In future using antibiotic resistance gene mRNAs as targets for ncRNA interaction may provide an optimal test for avoidance. Designing antibiotic resistant mRNAs with high affinity for ncRNA interactions in *E. coli* and culturing them in media supplemented with that antibiotic would provide strong selective pressure for improved resistance. This will allow us to test how quickly the cells can adapt by implementing avoidance strategies. Two common antibiotic resistance genes found in *E. coli* are the tetracycline resistance gene and the ampicillin resistance gene (Briñas, Zarazaga, Sáenz, Ruiz-Larrea, & Torres, 2002; Karami et al., 2006). Avoidance variants of these genes could quickly be generated using the same approach for the *araC* gene variants.

One problem with selecting *araC* mRNAs as a target for ncRNA interactions is that it prevented us from utilising a well-established method of strain identification in direct competition experiments. *E. coli* REL606 strains carry a selectively neutral *Ara*$^{-}$ mutation making them unable to metabolise arabinose. The *Ara*$^{-}$ mutation also has a second phenotypic effect when grown on TA agar plates which makes the cells appear pink or red. However, cells that are able to utilise arabinose will excrete acetic acid resulting in a change in the tetrazolium dye in

the agar from red to white (Remold & Lenski, 2001). The *E. coli* strain used in this study, REL607, is a derivative of REL606, with the only difference between them being a single point mutation in the *araA* gene that converts REL606 to an *Ara+* mutant, known as REL607. This mutation allows REL607 to metabolise arabinose and thus appears white on TA agar. Had a different gene been utilised in this study it would have been relatively easy to convert lines of REL607-SYN and REL607-INT to *Ara⁻* mutants essentially creating REL606-SYN and REL606-INT respectively and making them appear red on TA agar. The *Ara⁻* mutant strains could then be compared in a competition assay against the *Ara+* strains to determine whether one strain had a fitness advantage over the other. However, doing this would eliminate the cells ability to utilise arabinose, thus preventing any selective pressure being placed on the strains to improve their ability to metabolise arabinose.

Competition Experiments Comparing Avoidance Strains to REL606

The main finding from this research indicates that strains with an altered *araC* start site, which we predicted would increase interactions between the mRNA and ncRNAs in *E. coli* under conditions where arabinose is the sole carbon source in the media, did not exhibit altered growth rates. Due to time constraints, we were unable to assay the fitness of our engineered strains using direct competition experiments. Future experiments should test the relative fitness of the wild-type REL607 strain against the relative fitness of the designed REL607 strains (SYN and INT). Competition assays are designed to measure the relative fitness of one strain against another strain. While red/white screening via the *Ara⁻* mutation is not possible for our designed strains there are other methods that would allow us to distinguish between them. By adding selectively neutral opposite fluorescent markers each strain can be easily distinguished (Lind et al., 2017). Strains carrying such neutral selective markers already exist

as part of the Poole Lab strain collection. In future, it will be important to assess whether a general fitness difference exists between our designed REL607 strains and wild-type REL607. By measuring the net growth of two different populations, competitive fitness assays incorporate differences across the full culture cycle, which may include such fitness components as lag times, exponential growth rates, and stationary phase dynamics in batch culture (Wiser & Lenski, 2015). In the following paragraphs, I will outline a protocol for a competition assay that future researchers may implement to measure differences in fitness between the designed strains.

1) Revive strains (REL606, REL607::*araC*SYN and REL607::*araC*INT) from freezer stocks. Use a pipette tip to take a scraping of each strain and inoculate into 10ml LB. Incubate overnight culture at 37°C for 24 hours. Note that a standard O/N culture of E. coli in 5ml of LB can yield around 10^9 -10^{10} cells/ml (Sezonov, Joseleau-Petit, & D'Ari, 2007).

2) The following day dilute the two strains to be competed (REL606 vs REL607::*araC*SYN and REL606 vs REL607::*araC*INT) 200-fold and mix together in LB (50µl of each strain in 9.9ml of LB), this gives a 100-fold overall dilution in cell number (Leroi, Bennett, & Lenski, 1994; Wiser & Lenski, 2015).

3) Create six experimental replicates of each competition experiment (6 x REL606 vs REL607::*araC*SYN and 6 x REL606 vs REL607::*araC*INT). Immediately plate dilutions that yield 100-500 cells on tetrazolium and arabinose (TA) agar, this gives the initial frequencies of the two strains. The mixed strains can then be returned to the incubator to incubate at 37°C for 24 hours.

4) After exactly 24 hours create a serial dilution of each strain in each replicate experiment and plate final dilutions, 10^{-6} and 10^{-7}, on TA. To measure fitness more precisely continue to serially transfer for 3 days after initial plating.

During and immediately following the protocol outlined above colony forming unit counts should be taken. In microbiology, a colony-forming unit (CFU) is a unit used to estimate the number of viable bacteria or fungal cells in a sample (Sieuwerts, De Bok, Mols, De Vos, & Van Hylckama Vlieg, 2008). Viable is defined as an organism's ability to multiply via binary fission under controlled conditions. Counting with colony-forming units requires culturing the microbes and will thus count only viable cells, in contrast with absorbance measurements which counts all cells, living or dead. Thus, CFUs can be used to determine if there is a significant difference between the number of *viable* cells between the designed *araC* variant strains and REL606. In the following paragraphs, I will outline a CFU count protocol to take place during the 24-hour incubation period of the competition experiment.

At 3-hour intervals set up a series of dilution tubes to obtain dilutions of 10^{-1} through 10^{-7} of the *E. coli* strain cultures. Each dilution tube should contain 900ml of dilution fluid (minimal + arabinose). A dilution series is needed for each time interval from initial plating (T_0) through to the final plating (T_n). At each interval 36 TA plates are required, 1 plate per dilution (3 dilutions) per replicate (6 replicates) per strain (2 strains) (1 x 3 x 6 x 2 = 36), or using 12 selective plates have 3 count plates per test strain should be sufficient for resolving differences in mutation rates (Barrick Lab Protocols: http://barricklab.org) Create the serial dilution as follows:

1. Add 9.9ml of LB to 7 Eppendorf tubes labelled A through G.
2. Dilute mixed *E. coli* strains by adding 100ml from the tube labelled T1 to Tube A which contains 9.9ml LB. Tube A will be the 10^{-1} dilution of T_1.
3. Vortex 10^{-1} Tube for 5 seconds.
4. Following this add 100ml of Tube A (10^{-1}) to the next tube of LB (Tube B). Tube B is a 10^{-2} dilution of T1. Thus, each tube is a 10-fold dilution. Continue serially diluting in this manner until 100ml has been added to all tubes.
5. Plate dilutions that yield roughly 30-300 colonies, keep in mind this will change across time intervals. Pipette 100µl aliquots of the selected dilution tube to the center of the agar plate, spread using a flame sterilized glass rod. Repeat plating for each dilution series, T_1 through T_n cultures.
6. Once plates have dried, invert and incubate overnight at 37°C.

After the 24-hour incubation period measure colony counts using the ProtoCOL 3 Colony Counter (Synbiosis). CFU Counts can subsequently be used to determine the fitness of each strain in comparison to each other.

Applications of Avoidance

If the mRNA:ncRNA avoidance model is accurate there are several ways this knowledge can be applied to biological research. For instance, this research could be applied to the control and treatment of bacterial infections. Had time permitted, designing ncRNAs to bind to essential mRNA transcripts within the *E. coli* genome would have been an interesting application of this research. These ncRNAs would be targeted to genes that are highly

important for the growth and survival of *E. coli*; as such they will act as a type of antimicrobial agent that suppress the translation of these essential transcripts limiting the growth and overall fitness of *E. coli*. The effects of this treatment will be observed in vitro by designing 'sticky' ncRNAs and incorporating them into a high expression, high copy number plasmid to be transformed into *E. coli* cells, which simulates the natural stoichiometry of mRNA and ncRNA levels, as ncRNAs are available in much higher numbers. This will allow us to objectively determine any phenotypic impact these RNAs may have on the growth of the organism. This method also provides a quick assay for determining the essentiality of the target gene in the bacteria.

A possible setback of this method however is that the cells natural immune response may lead to degradation of the designed ncRNA (Quabius & Krupp, 2015). To circumvent this kind of activity it may be necessary to incorporate an Hfq protein-binding region into the design of the ncRNA. Hfq acts as a chaperone and is often found on stem loops of the RNAs to provide stability for the molecule and prevent degradation (Brennan & Link, 2007). The design for ncRNAs will focus on targeting the 5' end of the coding sequence (CDS) as it has been shown to be very important for the initiation of translation (Kudla et al., 2009; Plotkin & Kudla, 2011; Tuller & Zur, 2015), thus any disturbance in this region of the CDS will likely inhibit protein synthesis.

Alternative Methods for Introducing *araC* Variants into REL607

In hindsight, it may have been beneficial to use an alternative approach for introducing the variant copies of *araC* into REL607. The overall result of the gene replacement was a total of four nucleotide changes in the native *araC* gene for both *araC*SYN and *araC*INT. As only a few changes were made to a gene that already exists and is native to *E. coli* perhaps a better

approach would have been to use site-directed mutagenesis where the “knock-in” could have been performed using CRISPR/Cas to edit the *araC* gene in the chromosome of REL607. CRISPR allows the chromosomes of live cells to be edited at any location based on guide strand of RNA (Miller et al., 2017). The RNA-guided CRISPR associated protein 9 (Cas9) forms double stranded nicks in genomic DNA. Cas9 is guided by a RNA molecule, called a single guide RNA (sgRNA) that can be designed to target any site in the DNA. The Cas9:sgRNA complex recognizes the sequence complementary to the sgRNA. Following the double stranded break (DBS) of the DNA the DBS repair pathways can facilitate site-directed mutagenesis, insertions or deletions (Miller et al., 2017)

Concluding Remarks

The mRNA: ncRNA avoidance model has provided great new insights for the optimisation of precision bioengineering. Determining how such mechanisms have become established across species is important for understanding how evolution shapes RNA-RNA interactions. The results of this thesis highlight the complexity of this phenomenon. Understanding more about the nature of mRNA-ncRNA interactions and how they relate to protein expression presents an important undertaking in molecular biology.

References

Agrawal, R. K., Penczek, P., Grassucci, R. A., Li, Y., Leith, A., Nierhaus, K. H., & Frank, J. (1996). Direct Visualization of A-, P-, and E-Site Transfer RNAs in the Escherichia coli Ribosome . *Science, 271*(5251), 1000-1002.

Altman, S. (2011). Ribonuclease P. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1580), 2936–2941.

Ananth, P., Goldsmith, G., & Yathindra, N. (2013). An innate twist between Crick ' s wobble and Watson-Crick base pairs An innate twist between Crick ' s wobble and Watson-Crick base pairs, 1038–1053.

Andronescu, M. et al., 2014. The determination of RNA folding nearest neighbor parameters. *Methods in Molecular Biology,* 1097, 44-70.

Bandyra, K. J., Said, N., Pfeiffer, V., Górna, M. W., Vogel, J., & Luisi, B. F. (2012a). The Seed Region of a Small RNA Drives the Controlled Destruction of the Target mRNA by the Endoribonuclease RNase E. *Molecular Cell*, *47*(6), 943–953.

Bandyra, K. J., Said, N., Pfeiffer, V., Górna, M. W., Vogel, J., & Luisi, B. F. (2012b). The Seed Region of a Small RNA Drives the Controlled Destruction of the Target mRNA by the Endoribonuclease RNase E. *Molecular Cell*, *47*(6), 943–953.

Banerjee, S., Chalissery, J., Bandey, I., & Sen, R. (2006). Rho-dependent Transcription Termination: More Questions than Answers. *Journal of Microbiology (Seoul, Korea)*, *44*(1), 11–22.

Bao, M., Cervantes Cervantes, M., Zhong, L., & Wang, J. T. L. (2012). Searching for Non-coding RNAs in Genomic Sequences Using ncRNAscout. *Genomics, Proteomics & Bioinformatics*, *10*(2), 114–121.

Barrick, J. E., Yoon, S. H., Kim, J. F., Yu, Dong SuJeong, H., Oh, T. K., Lenski, R. E., & Schneider, D. (2009). Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature, 461*(7268, 1243-1247.

Beaumont, H. J. E., Gallie, J., Kost, C., Ferguson, G. C., & Rainey, P. B. (2009). Experimental evolution of bet hedging. *Nature*, *462*(7269), 90–93.

Blount, Z. D., Borland, C. Z., & Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(23), 7899–7906.

Boël, G., Letso, R., Neely, H., Price, W. N., Wong, K.-H., Su, M., … Hunt, J. F. (2016). Codon influence on protein expression in E. coli correlates with mRNA levels . *Nature,*

529(7586), 385-363.

Bornhorst, J. A., & Falke, J. J. (2000). [16] Purification of Proteins Using Polyhistidine Affinity Tags. *Methods in Enzymology*, *326*, 245–254.

Bradshaw, N., & Walter, P. (2007). The Signal Recognition Particle (SRP) RNA Links Conformational Changes in the SRP to Protein Targeting. *Molecular Biology of the Cell*, *18*(7), 2728–2734.

Brennan, R. G., & Link, T. M. (2007). Hfq structure, function and ligand binding. *Current Opinion in Microbiology*, *10*(2), 125–133.

Bryksin, A. V, & Matsumura, I. (2010). Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. *BioTechniques*, *48*(6), 463–465.

Carpenter, S., Ricci, E. P., Mercier, B. C., Moore, M. J., & Fitzgerald, K. A. (2014). Post-transcriptional regulation of gene expression in innate immunity . *Nature Reviews. Immunology, 14*(6), 361-376.

Carthew, R. W., & Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*, *136*(4), 642–655.

Clark, M. B., Amaral, P. P., Schlesinger, F. J., Dinger, M. E., Taft, R. J., Rinn, J. L., … Mattick, J. S. (2011). The reality of pervasive transcription . *PLoS Biology, 9*(7), e1000625.

Cloonan, N. (2015). Re-thinking miRNA-mRNA interactions: Intertwining issues confound target discovery. *BioEssays*, *37*(4), 379–388.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423.

Consortium, T. E. P. (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature*, *489*(7414), 57–74.

Dana, A., & Tuller, T. (2014). The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Research*, *42*(14), 9171–9181.

Darty, K., Denise, A., & Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, *25*(15), 1974–1975.

Débarbouillé, M., Gabay, J., Schwartz, M., & Hall, M. N. (1982). A role for mRNA secondary structure in the control of translation initiation . *Nature, 295*(5850), 616-618.

Desai, T. A., & Rao, C. V. (2010). Regulation of Arabinose and Xylose Metabolism in Escherichia coli . *Applied and Environmental Microbiology*, *76*(5), 1524–1532.

Deutscher, M. P. (2006). Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic Acids Research*, *34*(2), 659–666.

Dieterich, C., & Stadler, P. F. (2013). Computational biology of RNA interactions. *Wiley Interdisciplinary Reviews: RNA*, *4*(1), 107–120.

Dirla, S., Chien, J. Y.-H., & Schleif, R. (2009). Constitutive Mutations in the Escherichia coli AraC Protein . *Journal of Bacteriology*, *191*(8), 2668–2674.

Diwa, A., Bricker, A. L., Jain, C., & Belasco, J. G. (2000). An evolutionarily conserved RNA stem–loop functions as a sensor that directs feedback regulation of RNase E gene expression. *Genes & Development*, *14*(10), 1249–1260.

Dodd, D. M. B. (1989). Reproductive Isolation as a Consequence of Adaptive Divergence in Drosophila pseudoobscura. *Evolution*, *43*(6), 1308–1311.

Ellis, R. J. (2001). Macromolecular crowding: obvious but underappreciated . *Trends in Biochemical Sciences, 26*(10), 597-604.

Englesberg, E. (1961). ENZYMATIC CHARACTERIZATION OF 17 l-ARABINOSE NEGATIVE MUTANTS OF ESCHERICHIA COLI. *Journal of Bacteriology*, *81*(6), 996–1006.

Fehér, T., Karcagi, I., Gyorfy, Z., Umenhoffer, K., Csörgo, B., & Pósfai, G. (2008). Scarless engineering of the Escherichia coli genome. *Methods in Molecular Biology (Clifton, N.J.)*, *416*(2), 251–259.

Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., … Bateman, A. (2009). Rfam: updates to the RNA families database. *Nucleic Acids Research*, *37*(Supplement 1), D136–D140.

Giannoukos, G., Ciulla, D. M., Huang, K., Haas, B. J., Izard, J., Levin, J. Z., … Gnirke, A. (2012). Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biology*, *13*(3), r23.

Gomes, A. Q., Nolasco, S., & Soares, H. (2013). Non-Coding RNAs: Multi-Tasking Molecules in the Cell. *International Journal of Molecular Sciences*, *14*(8), 16010–16039.

Gottesman, S., & Storz, G. (2011). Bacterial Small RNA Regulators: Versatile Roles and Rapidly Evolving Variations. *Cold Spring Harbor Perspectives in Biology*, *3*(12), a003798.

Gu, W., Zhou, T., & Wilke, C. O. (2010). A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes. *PLoS Computational Biology*, *6*(2), e1000664.

Gustafsson, C., Govindarajan, S., & Minshull, J. (2004). Codon bias and heterologous protein expression. *Trends in Biotechnology*, *22*(7), 346–353.

Hall, B. G., Acar, H., Nandipati, A., & Barlow, M. (2014). Growth rates made easy . *Molecular*

Biology and Evolution, 31(1), 232-238.

Hausser, J., & Zavolan, M. (2014). Identification and consequences of miRNA-target interactions -- beyond repression of gene expression. *Nature Reviews. Genetics*, *15*(10), 702.

Herbig, A., & Nieselt, K. (2011). nocoRNAc: Characterization of non-coding RNAs in prokaryotes. *BMC Bioinformatics*, *12*(1), 40.

Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of Molecular Biology*, *151*(3), 389–409.

Kastritis, P. L., & Bonvin, A. M. J. J. (2013). On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of the Royal Society Interface*, *10*(79), 20120835.

Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I., & Whitlock, M. C. (2012). Experimental evolution. *Trends in Ecology & Evolution*, *27*(10), 547–560.

Keiler, K. C., & Ramadoss, N. S. (2011). Bifunctional transfer-messenger RNA. *Biochimie*, *93*(11), 1993–1997.

Kimple, M. E., Brill, A. L., & Pasker, R. L. (2013). Overview of Affinity Tags for Protein Purification. *Current Protocols in Protein Science, 73*, Unit-9.9.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., & Haussler, D. (1994). Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *Journal of Molecular Biology*, *235*(5), 1501–1531.

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. *Science (New York, N.Y.)*, *324*(5924), 255–258.

Kuznetsova, I. M., Turoverov, K. K., & Uversky, V. N. (2014). What Macromolecular Crowding Can Do to a Protein. *International Journal of Molecular Sciences*, *15*(12), 23090–23140.

Lai, D., & Meyer, I. M. (2016). A comprehensive comparison of general RNA-RNA interaction prediction methods. *Nucleic Acids Research*, *44*(7), e61–e61.

Lang, F. (2007). Mechanisms and Significance of Cell Volume Regulation . *Journal of the American College of Nutrition, 26*(Supplement 5), 613S.

Lenski, R. E. (1988). Experimental Studies of Pleiotropy and Epistasis in Escherichia coli. I. Variation in Competitive Fintess Among Mutants Resistant to Virus T4. *Evolution*, *42*(3), 425–432.

Lenski, R. E., Rose, M. R., Simpson, S. C., & Tadler, S. C. (1991). Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations . *The American Naturalist, 138*(6), 1315-1341.

Leroi, A. M., Bennett, A. F., & Lenski, R. E. (1994). Temperature acclimation and competitive fitness: an experimental test of the beneficial acclimation assumption. *Proceedings of the National Academy of Sciences* , *91*(5), 1917–1921.

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*, *120*(1), 15–20.

Licon, A., Taufer, M., Leung, M.-Y., & Johnson, K. L. (2010). A Dynamic Programming Algorithm for Finding the Optimal Segmentation of an RNA Sequence in Secondary Structure Predictions. *2nd International Conference on Bioinformatics and Computational Biology 2010, (BICoB-2010), Honolulu, Hawaii, USA, 24-26 March 2010. International Conference on Bioinformatics and Computational Biology (2nd : 2010 : Honolulu, Hawaii)*, *2010*, 165–170.

Lind, P. A., Arvidsson, L., Berg, O. G., & Andersson, D. I. (2017). Variation in Mutational Robustness between Different Proteins and the Predictability of Fitness Effects. *Molecular Biology and Evolution*, *34*(2), 408–418.

Lindgreen, S., Umu, S. U., Lai, A. S.-W., Eldai, H., Liu, W., McGimpsey, S., … Gardner, P. P. (2014). Robust Identification of Noncoding RNA from Transcriptomes Requires Phylogenetically-Informed Sampling. *PLoS Computational Biology*, *10*(10), e1003907.

Lithwick, G., & Margalit, H. (2003). Hierarchy of sequence-dependent features associated with prokaryotic translation . *Genome Research, 13*(12), 2665-2673.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26.

Lorenz, R., Wolfinger, M. T., Tanzer, A., & Hofacker, I. L. (2016). Predicting RNA secondary structures from sequence and probing data. *Methods*, *103*, 86–98.

Maier, T., Güell, M., & Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples . *FEBS Letters, 583*(24), 3966-3973.

Maloy, S. R., Stewart, V. J., & Taylor, R. K. (1996). Genetic analysis of pathogenic bacteria: a laboratory manual . Plainview, N.Y : Cold Spring Harbor Laboratory Press

Mao, Y., Liu, H., Liu, Y., & Tao, S. (2014). Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in Saccharomyces cerevisiae. *Nucleic Acids Research*, *42*(8), 4813–4822.

Marlatt, N. M., Spratt, D. E., & Shaw, G. S. (2010). Codon optimization for enhanced Escherichia coli expression of human S100A11 and S100A1 proteins. *Protein Expression and Purification*, *73*(1), 58–64.

Miller, J. B., Zhang, S., Kos, P., Xiong, H., Zhou, K., Perelman, S. S., … Siegwart, D. J. (2017). Non-viral CRISPR/Cas gene editing in vitro and in vivo enabled by synthetic nanoparticle co-delivery of Cas9 mRNA and sgRNA. *Angewandte Chemie (International Ed. in English)*, *56*(4), 1059–1063.

Milo, R. (2013). What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays*, *35*(12), 1050–1055.

Mourão, M. A., Hakim, J. B., & Schnell, S. (2014). Connecting the Dots: The Effects of Macromolecular Crowding on Cell Physiology. *Biophysical Journal*, *107*(12), 2761–2766.

Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F., & Hofacker, I. L. (2006a). Thermodynamics of RNA–RNA binding. *Bioinformatics*, *22*(10), 1177–1182.

Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F., & Hofacker, I. L. (2006b). Thermodynamics of RNA–RNA binding. *Bioinformatics*, *22*(10), 1177–1182.

Onoa, B., & Tinoco, I. (2004). RNA folding and unfolding. *Current Opinion in Structural Biology*, *14*(3), 374–379.

Pain, A., Ott, A., Amine, H., Rochat, T., Bouloc, P., & Gautheret, D. (2015). An assessment of bacterial small RNA target prediction programs. *RNA Biology*, *12*(5), 509–513.

Papenfort, K., Bouvier, M., Mika, F., Sharma, C. M., & Vogel, J. (2010). Evidence for an autonomous 5′ target recognition domain in an Hfq-associated small RNA. *Proceedings of the National Academy of Sciences*, *107*(47), 20435 LP-20440.

Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. TL - 12. *Nature Reviews. Genetics*, *12 VN-r*(1), 32–42.

Pósfai, G., Koob, M. D., Kirkpatrick, H. A., & Blattner, F. R. (1997). Versatile insertion plasmids for targeted genome manipulations in bacteria: isolation, deletion, and rescue of the pathogenicity island LEE of the Escherichia coli O157:H7 genome. *Journal of Bacteriology*, *179*(13), 4426–4428.

Proutski, V., & Holmes, E. (1998). SWAN: sliding window analysis of nucleotide sequence variability. *Bioinformatics*, *14*(5), 467–468.

Qian, W., He, X., Chan, E., Xu, H., & Zhang, J. (2011). Measuring the evolutionary rate of protein–protein interaction. *Proceedings of the National Academy of Sciences* , *108*(21), 8725–8730.

Qian, W., & Zhang, J. (2014). Genomic evidence for adaptation by gene duplication. *Genome Research*, *24*(8), 1356–1362.

Quabius, E. S., & Krupp, G. (2015). Synthetic mRNAs for manipulating cellular phenotypes: an overview. *New Biotechnology*, *32*(1), 229–235.

Quax, T. E. F., Claassens, N. J., Söll, D., & van der Oost, J. (2015). Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell*, *59*(2), 149–161.

Ralston, G. B. (1990). Effects of "crowding" in protein solutions . *Journal of Chemical Education, 67*(10), 857.

Remold, S. K., & Lenski, R. E. (2001). Contribution of individual random mutations to genotype-by-environment interactions in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11388–11393.

Rice, W. R. (1996). Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature*, *381*(6579), 232–234.

Salari, R., Backofen, R., & Sahinalp, S. C. (2010). Fast prediction of RNA-RNA interaction. *Algorithms for Molecular Biology : AMB*, *5*, 5.

Sambrook, J., Russell, D. W., & Maniatis, T. (2001). Molecular cloning: a laboratory manual . Cold Spring Harbor, N.Y : Cold Spring Harbor Laboratory Press.

Sargo, C., Campani, G., Silva, G., Correia, D., Giordano, R. de, Ferreira, E., … Zangirolami, T. (2015). Salmonella typhimurium and Escherichia coli dissimilarity: Closely related bacteria with distinct metabolic profiles. *Biotechnology progress*, *31*(5), 1217-1225.

Schleif, R. (2010). AraC protein, regulation of the l-arabinose operon in Escherichia coli, and the light switch mechanism of AraC action. *FEMS Microbiology Reviews*, *34*(5), 779–796.

Sergiev, P. V, Lesnyak, D. V, Kiparisov, S. V, Burakovsky, D. E., Leonov, A. A., Bogdanov, A. A., … Dontsova, O. A. (2005). Function of the ribosomal E-site: a mutagenesis study. *Nucleic Acids Research*, *33*(18), 6048–6056.

Sezonov, G., Joseleau-Petit, D., & D'Ari, R. (2007). Escherichia coli Physiology in Luria-Bertani Broth . *Journal of Bacteriology*, *189*(23), 8746–8749.

Sharp, P. M., & Li, W. H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, *15*(3), 1281–1295.

Shimomura, O. (2005). The discovery of aequorin and green fluorescent protein. *Journal of Microscopy*, *217*(1), 3–15.

Siegele, D. A., & Hu, J. C. (1997). Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(15), 8168–8172.

Sieuwerts, S., De Bok, F. A. M., Mols, E., De Vos, W. M., & Van Hylckama Vlieg, J. E. T. (2008). A simple and fast method for determining colony forming units. *Letters in Applied Microbiology*, *47*(4), 275–278.

Singh, S. S., Singh, N., Bonocora, R. P., Fitzgerald, D. M., Wade, J. T., & Grainger, D. C. (2014). Widespread suppression of intragenic transcription initiation by H-NS. *Genes & Development*, *28*(3), 214–219.

Sørensen, M. A., Kurland, C. G., & Pedersen, S. (1989). Codon usage determines translation rate in Escherichia coli. *Journal of Molecular Biology*, *207*(2), 365–377.

Steuten, B., Hoch, P. G., Damm, K., Schneider, S., Köhler, K., Wagner, R., & Hartmann, R. K. (2014). Regulation of transcription by 6S RNAs: Insights from the Escherichia coli and Bacillus subtilis model systems. *RNA Biology*, *11*(5), 508–521.

Stoner, C. M., & Schleif, R. (1982). Is the amino acid but not the nucleotide sequence of the Escherichia coli araC gene conserved? *Journal of Molecular Biology*, *154*(4), 649–652.

Storz, G., Vogel, J., & Wassarman, K. M. (2011). Regulation by Small RNAs in Bacteria: Expanding Frontiers. *Molecular Cell*, *43*(6), 880–891.

Studer, S. M., & Joseph, S. (2006). Unfolding of mRNA Secondary Structure by the Bacterial Translation Initiation Complex. *Molecular Cell*, *22*(1), 105–115.

Stülke, J., & Hillen, W. (1999). Carbon catabolite repression in bacteria. *Current Opinion in Microbiology*, *2*(2), 195–201.

Tamai, E., Belyaeva, T. A., Busby, S. J. W., & Tsuchiya, T. (2000). *Mutations That Increase the Activity of the Promoter of the Escherichia coli Melibiose Operon Improve the Binding of MelR, a Transcription Activator Triggered by Melibiose*. *The Journal of biological chemistry, 275*(22),17058-17063

Tuller, T., Waldman, Y. Y., Kupiec, M., Ruppin, E., & Sherman, F. (2010). Translation Efficiency Is Determined by Both Codon Bias and Folding Energy . *Proceedings of the National Academy of Sciences of the United States of America, 107*(8), 3645-3650.

Tuller, T., & Zur, H. (2015). Multiple roles of the coding sequence 5’ end in gene expression regulation . *Nucleic Acids Research, 43*(1), 13-28.

Udekwu, K. I. (2010). Transcriptional and Post-Transcriptional Regulation of the Escherichia coli luxS mRNA; Involvement of the sRNA MicA. *PLOS ONE*, *5*(10), e13449.

Umu, S. U., & Gardner, P. P. (2017). A comprehensive benchmark of RNA–RNA interaction prediction tools for all domains of life. *Bioinformatics*, *33*(7), 988–996.

Umu, S. U., Poole, A. M., Dobson, R. C. J., & Gardner, P. P. (2016). Avoidance of stochastic RNA interactions can be harnessed to control protein expression levels in bacteria and archaea. *eLife*, *5*, e13479.

Uzman, A. (2001). Molecular Cell Biology (4th edition) Harvey Lodish, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, David Baltimore and James Darnell; Freeman & Co., New York, NY, 2000, 1084, ISBN 0-7167-3136-3 . *Biochemistry and Molecular Biology Education* .

Vogel, J., & Luisi, B. F. (2011). Hfq and its constellation of RNA. *Nature Reviews. Microbiology*, *9*(8), 578–589.

Vogel, J., & Wagner, E. G. H. (2007). Target identification of small noncoding RNAs in bacteria. *Current Opinion in Microbiology*, *10*(3), 262–270.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63.

Wen, J.-D., Lancaster, L., Hodges, C., Zeri, A.-C., Yoshimura, S. H., Noller, H. F., … Tinoco, I. (2008). Following translation by single ribosomes one codon at a time. *Nature*, *452*(7187), 598–603.

Westhof, E., & Fritsch, V. (2000). RNA folding: beyond Watson–Crick pairs. *Structure*, *8*(3), R55–R65.

Wiser, M. J., & Lenski, R. E. (2015). A Comparison of Methods to Measure Fitness in Escherichia coli. *PLoS ONE*, *10*(5), e0126210.

Wiser, M. J., Ribeck, N., & Lenski, R. E. (2013). Long-term dynamics of adaptation in asexual populations . *Science, 342*(6164) 1364-1367.

Wu, Y., Wei, B., Liu, H., Li, T., & Rayner, S. (2011). MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC Bioinformatics*, *12*, 107.

Yang, J.-R., Liao, B.-Y., Zhuang, S.-M., & Zhang, J. (2012). Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proceedings of the National Academy of Sciences* , *109*(14), E831–E840.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, *4*(11), 218.

Zhao, Y., Shen, X., Tang, T., & Wu, C.-I. (2017). Weak Regulation of Many Targets Is Cumulatively Powerful—An Evolutionary Perspective on microRNA Functionality.

Molecular Biology and Evolution, *34*(12), 3041–3046.

Zhong, S., Joung, J. G., Zheng, Y., Chen, Y. R., Liu, B., Shao, Y., … Giovannoni, J. J. (2011). High-throughput illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb Protocols*, *2011*.

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule . *Science, 244*(4900), 48-52

Zuker, M., & Sankoff, D. (1984). RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, *46*(4), 591–621.